MEASUREMENT DRIVEN INSTRUCTION: A LONGITUDINAL STUDY

By

GAYLE R. McLAURIN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF EDUCATION

UNIVERSITY OF FLORIDA

1993

UNIVERSITY OF FLORIDA LIBRARIES

ACKNOWLEDGMENTS

I wish to thank Dr. Linda Crocker for her assistance and constant encouragement throughout my doctoral program. The completion of this degree program as well as this research endeavor were made bearable because Dr. Crocker exhibits such a sense of caring for students and a sense of professionalism for educational research. Dr. David Miller also provided invaluable technical assistance and understanding for the problems encountered during this research. I would also like to extend appreciation to Dr. Craig Frisby and to Dr. Craig Wood as committee members for their editorial comments. Finally, I would have been unable to complete this program without the patience and encouragement from my husband, Mark, and sons, Joshua and Adam.

Words never seem adequate when one attempts to express gratitude, but it is hoped those previously mentioned will understand the depth of feeling when I use the words "thank you."

This research was funded, in part, by the Florida Education Research Council whose executive director is Dr. Charlie Council.

TABLE OF CONTENTS

		Page
ACKNOWLEDGMENTS		ii
LIST OF TABLES		v
ABSTRACT		ix
CHAPTERS		
1	INTRODUCTION	1
	Concerns About Measurement Driven Instruction	2
	Description of this Study	5
	Significance of the Study	8
	Overview of the Dissertation	9
2	REVIEW OF LITERATURE	11
	Curricular Alignment: Effects on Instruction	11
	Empirical Studies of Curricular Alignment: Effects on Test Results	26
	Curricular Sensitivity and Dimensionality of Achievement Tests	29
	Summary	41
3	METHODOLOGY	44
	Study Setting	44
	Sample	45
	Instruments	46
	Design and Data Collection	48
	Method of Analysis	49
4	RESULTS	53
	Research Questions	53
	Summary	97

5 DISCUSSION AND IMPLICATIONS FOR FURTHER RESEARCH. 99

Summary of Findings..... 99
Limitations..... 109
Implications for Further Research..... 111

APPENDICES

A TABLES RESULTING FROM FACTOR ANALYSIS OF SUBSKILL AREAS..... 113

B FIRST 10 EIGENVALUES..... 122

C CORRELATIONS AMONG TEST PREPARATION VARIABLES... 126

REFERENCES..... 132

BIOGRAPHICAL SKETCH..... 138

LIST OF TABLES

Table		Page
1	Teacher Expectation of Student Testing with Testing Program Implemented in 1987.....	6
2	Number of Teachers and Students in Sample by Year and Grade.....	46
3	Average Number of Test Preparation Activities in by Grade and Year.....	56
4	Significant Differences in Test Preparation Activities during Year 3 between Pairs of Grade Levels.....	57
5	Grade 2 Comparison of Mean Raw Scores by Year....	60
6	Grade 3 Comparison of Mean Raw Scores by Year....	60
7	Grade 4 Comparison of Mean Raw Scores by Year....	61
8	Grade 5 Comparison of Mean Raw Scores by Year....	61
9	Mean Raw Score (Standard Deviation) for Each Grade and Year on Mathematics Subtest, CAT.....	62
10	Math Computation: Numbers of Factors by Grade and Year.....	66
11	Math Concepts and Applications: Numbers of Factors by Grade and Year.....	66
12	Grade 2 Mathematics Computation Factor Loadings of Items for 3 Years.....	69
13	Grade 2 Mathematics Concepts and Applications: Item Factor Loadings for 3 Years.....	70
14	Grade 2 Goodness-of-Fit Index for Each Year and Subtest.....	72
15	Grade 3 Mathematics Computation Item Factor Loadings for 3 Years based on Year 1.....	74

16 Grade 3 Goodness-of-Fit Index for Each Year and Subtest..... 76

17 Grade 3 Concepts and Applications Item Factor Loadings for 3 Years..... 77

18 Grade 3 Mathematics Concepts and Applications Skill-Area Subtest Factor Loadings for 3 Years... 79

19 Grade 3 Skill-Area Subtests Goodness-of-Fit Index for Each Year and Subtest..... 80

20 Grade 4 Mathematics Computation Item Factor Loadings for 3 Years..... 82

21 Goodness-of-Fit Index for Each Year and Subtest.. 83

22 Grade 4 Mathematics Computation Skill-Area Subtest Factor Loadings for 3 Years..... 84

23 Grade 5 Mathematics Computation Item Factor Loadings for 3 Years..... 88

24 Grade 5 Goodness-of-Fit Index for Each Subtest... 90

25 Grade 5 Mathematics Concepts and Applications Item Factor Loadings for 3 Years..... 91

26 Grade 5 Mathematics Concepts and Applications Skill-Area Subtest Factor Loadings for 3 Years... 93

27 Grade 5 Goodness-of-Fit Index for Skill-Area Subtests..... 94

28 Goodness-of-Fit Index Resulting from Applying Year 2 Factor Structures to Data from All Years.. 95

29 Goodness-of-Fit Index Resulting from Applying Year 3 Factor Structures to Data from All Years.. 96

30 Changes in Average Raw Score Means Expressed as Percentile Ranks..... 104

A-1 Grade 2 Mathematics Computation Skill-Area Subtest Factor Loadings for 3 Years..... 113

A-2 Grade 2 Skill-Area Subtest Goodness-of-Fit Index for Each Year and Subtest..... 113

A-3 Grade 2 Mathematics Concepts and Applications Skill-Area Subtest Factor Loadings for 3 Years... 114

A-4 Grade 2 Mathematics Computation Skill-Area Subtest Factor Loadings for 3 Years..... 115

A-5 Grade 4 Mathematics Computation Skill-Area Subtest Factor Loadings for 3 Years..... 116

A-6 Grade 4 Skill-Area Subtests Goodness-of-Fit Index for Each Year..... 117

A-7 Grade 4 Mathematics Concepts and Applications Skill-Area Subtest Factor Loadings by Year..... 117

A-8 Grade 5 Mathematics Computation Skill-Area Factor Loadings for 3 Years..... 118

A-9 Skill-Area Subtest Goodness-of-Fit Index Resulting from Application of Year 2 Factor Structures from Data of All Years..... 119

A-10 Skill-Area Subtest Goodness-of-Fit Index Resulting from Application of Year 3 Factor Structures from Data of All Years..... 120

B-1 First 10 Eigenvalues Resulting from Factor Analyses--1987..... 122

B-2 First 10 Eigenvalues Resulting from Factor Analyses--1988..... 123

B-3 First 10 Eigenvalues Resulting from Factor Analyses--1989..... 124

C-1 Correlations for Grade 2, 1987..... 126

C-2 Correlations for Grade 2, 1988..... 126

C-3 Correlations for Grade 2, 1989..... 127

C-4 Correlations for Grade 3, 1987..... 127

C-5 Correlations for Grade 3, 1988..... 128

C-6 Correlations for Grade 3, 1989..... 128

C-7 Correlations for Grade 4, 1987..... 129

C-8 Correlations for Grade 4, 1988..... 129

C-9 Correlations for Grade 4, 1989..... 130
C-10 Correlations for Grade 5, 1987..... 130
C-11 Correlations for Grade 5, 1988..... 131
C-12 Correlations for Grade 5, 1989..... 131

Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Education

MEASUREMENT DRIVEN INSTRUCTION:
A LONGITUDINAL STUDY

By

Gayle R. McLaurin

May 1993

Chair: Linda Crocker
Major Department: Foundations of Education

This study was designed to provide a description of the extent to which test preparation and curricular alignment activities took place over a 3-year period in a middle-sized metropolitan school district in north-central Florida. The focal years of the study encompassed the first year of the implementation of a nationally normed, standardized achievement test through the third year of its use. The classes of teachers who taught the same grade level for the 3 years in question served as subjects. Data on both teacher behavior and student test performance were collected. Impact of teacher behavior on student performance at item, subskill area, and total test levels for mathematics computation and mathematics concepts were examined. Evidence was presented that indicated that teachers change test preparation activities based on beliefs

about whether students would be tested and, to some degree, on the perceived emphasis placed on testing by the school board. Relationships were found between some test preparation activities in the classroom and average scores for the class. Average raw scores increased from the first to second year and then decreased in the third year of testing. Factor structures changed in that the loadings of some items on the varying factors changed from one year to the next.

CHAPTER 1
INTRODUCTION

Since the mid-1960s, interest in student achievement outcomes has been of central concern for many educators. This increasing interest in student testing was a direct result of the demand for education reform that originated with the Coleman report of 1966. Politicians and the public who elected them have pressed for accountability in education, and those efforts have resulted in a booming testing-related industry. According to Baker and Herman (1983), "Public concern about the effectiveness of schools has led to a reliance on testing and test results that is unprecedented in recent educational history" (p. 149). Because direct reform of educational practices in the classroom was impossible to implement from the offices of policymakers, those policymakers "made policy about testing, an available, well-developed relatively cheap and administratively simple technology" (p. 108). The result has been that, for many educators, tests have become "coercive device[s] to influence both the curriculum and instruction" (p. 108). As test results move to the forefront of national educational attention, those objectives that are being tested have greater impact on

instruction. The adaptation of classroom instructional practice to standardized test content and format has become known as curriculum alignment or measurement driven instruction (Brandt, 1989; Popham, 1987).

Concerns About Measurement Driven Instruction

Linn (1989) indicated that a tight link between test and curriculum objectives results in problems with respect to generalizability and the validity of achievement test scores. Those problems that may result from this increasing trend include the following:

1. Narrowing of the curriculum (Madaus, 1983; Shepard, 1989, 1990; Smith, 1991);
2. Overemphasis on skills that are readily tested at the expense of more complex thinking skills (Darling-Hammond & Wise, 1985; Frederiksen, 1984; Haertel & Calfee, 1983);
3. Reduced instructional time in subject areas as class time is devoted to practice activities for testing (Brandt, 1989; Mehrens & Kaminski, 1989; Nolen, Haladyna, & Haas, 1992);
4. "Test score pollution," which refers to increases or decreases in test performance that are unrelated to changes in the construct purportedly measured by the test (Haladyna, Nolen, & Haas, 1991; Nolen, Haladyna, & Haas, 1992; Shepard, 1990).

Although these concerns are serious, there is no uniform agreement among experts that curricular alignment activities are taken seriously enough by teachers to have significant impact on standardized test results. Researchers who have found differences in test or item performance that have been attributed to curricular differences have typically studied national samples drawn from a wide spectrum of school districts. Such data sets have substantial variance in both student characteristics and curricular offerings (e.g., Cook, Eignor, & Taft, 1988; Miller & Linn, 1988; Muthen, Kao, & Burstein, 1991). Consequently, Linn (1989) has suggested that very large differences in instructional offerings may be required to detect substantial effects on test performance. Such major differences in instructional offerings are more likely due to causes other than curricular alignment to the standardized test used in a district.

When the problem has been investigated at the local district level, studies of influence of curricular alignment on test performance have been less conclusive. After conducting a series of studies of total test performance, item parameter estimates, and factor structures of achievement test data, Mehrens and Phillips (1987) concluded that "the naturally occurring differences in curricula within a school district are not great enough to result in observable differences on standardized tests" (p. 368).

Unfortunately, Mehrens and Phillips failed to report how long the districts in their studies had been using the same standardized achievement test. If the same test had been in place for several consecutive years, substantial adjustments in classroom practice could already have occurred. In contrast to the findings of Mehrens and Phillips, researchers who have studied teacher reactions through classroom ethnographic methods (Smith, 1991) or surveys (Haladyna, Nolen, & Haas, 1991) have reported that teachers may alter their instruction dramatically to accommodate the content of district standardized achievement tests. So far, however, these researchers have not attempted to establish the link of teacher perceptions and actions to the subsequent test performance of their students. In one of the most promising efforts to date, Koretz, Linn, Dunbar, and Shepard (1991) have attempted to assess the effects of instructional alignment within a single district on a standardized achievement test of mathematics. They concluded that students were prepared for testing in ways that "boost scores . . . more than actual achievement" (p. 2), but this interpretation is somewhat clouded by the fact that different editions of the standardized achievement test were used, and, thus, their results could be explained by differential fit between the tests and the mathematics curriculum. It would be only natural to expect a district's

currently used test to fit its current published curriculum better than an earlier edition of that test series.

Description of this Study

Purpose

The purpose of this study was to document and describe the extent to which test preparation and curricular alignment activities took place over a 3-year period in a middle-sized southeastern metropolitan school district and to assess the effects of such curricular alignment on test performance and factor structure of a norm-referenced standardized test of mathematics achievement. In contrast to the studies of Mehrens and Phillips (1986, 1987), the focal years of longitudinal study encompass the first year of the implementation of a new standardized achievement test through the third year of its use. A further refinement is that only classes of teachers who taught the same grade level for the years in question were used as subjects. This assured that teachers had increasing familiarity with the test over time. Finally, this district provided a unique opportunity to expand the study of longitudinal adaptation of teachers and their students because school board action created a situation where approximately half of the teachers did not expect to administer the standardized achievement test to their students during the third year. Specifically, second- and fifth-grade teachers expected their students to be tested in all 3 years of the study, but third- and

fourth-grade teachers did not expect their students to be tested in the third year of study (see Table 1). This provided an unprecedented opportunity for comparison of performance of students at selected grade levels over a 3-year period in years when teachers did and then did not expect their students to be tested.

Table 1
Teacher Expectation of Student Testing with Testing Program Implemented in 1987

Grade	1987	1988	1989
2	yes	yes	yes
3	yes	yes	no
4	yes	yes	no
5	yes	yes	yes

Both teacher survey responses and student test results were examined in order to determine the impact of test preparation and curricular alignment activities in this district. Effects of teacher activities on standardized achievement test performance were assessed in terms of overall student test performance and the effects on the factor structures of students' dichotomous item responses and skill-area subtest scores as well.

Research Questions

The specific research questions addressed were as follows:

1. How did test preparation activities change over a 3-year period following adoption of a new achievement test? Did the number of activities differ by grade level depending on whether there was an expectation to test in the third year?
2. Was overall student test performance in mathematics affected by teacher curricular alignment or test-preparation activities?
3. At second and fifth grades, where teachers expected to be tested each year, were there increases in student math test scores over the 3-year period?
4. At third and fourth grades, was there a decrease in student math test scores in the year when teachers did not expect to be tested?
5. For each grade level, did the factor structure of dichotomous item responses on the mathematics test remain stable over the 3-year period in terms of number of factors, similarity in pattern of item loadings, and magnitude of goodness-of-fit indexes?
6. For each grade level, did the factor structure of the skill-area subtest scores remain stable over

the 3-year period (in terms of number of factors, similarity in pattern of subtest loading on factors, and magnitude of goodness-of-fit indexes)?

The analytic procedures used to answer these questions include descriptive statistics, multiple regression, analysis of variance, exploratory factor analysis, and confirmatory factor analysis. Analyses were conducted separately for each grade level.

Significance of the Study

As test results move to the forefront of national educational attention, tested objectives tend to have a great impact on instruction. This may lead to problems of narrowing and trivializing of curriculum and reduced instructional time in academic subjects. Prior to the present study, however, efforts to document the impact of curricular alignment on level of test performance or underlying test dimensionality have yielded contradictory findings, but this may have been due to the researchers' procedures. No study to date has been conducted with control for test forms, teacher pool, initial-plus-consecutive-year test use, and variation in teacher expectations about test administration. If differential curricular alignment and test preparation practices have demonstrable impact on standardized achievement test results over time when previous methodological limitations are controlled, this may

call into question many of the proposed and current uses and interpretations of standardized achievement test results.

Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). He emphasized that the use of test scores (i.e., interpretations) are to be validated rather than the tests themselves. Messick suggested that score meanings can be examined best by studying responses across time, in different settings, and under different conditions. It is within this framework that the research questions of the current study were posed.

Overview of the Dissertation

This study was designed to provide a description of the extent to which test preparation and curricular alignment activities took place over a 3-year period in a middle-sized metropolitan school district in north-central Florida. The focal years of the study encompassed the first year of the implementation of a new standardized achievement test through the third year of its use. The classes of teachers who taught the same grade level for the 3 years in question served as subjects. Data on both teacher behavior and student test performance were collected. Impact of teacher behavior on student performance at item, subskill area, and total test levels were examined.

Further discussion of empirical research regarding curricular alignment and methods of data analysis is presented in the review of literature, Chapter 2. In Chapter 3, a description of the methodology for conducting this research study can be found. The results of the data analysis will be presented in Chapter 4. A discussion of the results and implications for further study will be presented in Chapter 5.

CHAPTER 2
REVIEW OF LITERATURE

Several bodies of conceptual and empirical literature provide the background for investigation of the effects of curricular alignment. One relevant set of literature deals with writings that specifically address the issues of curricular alignment and its effects on classroom instructional practice. A second strand of relevant literature focuses on empirical attempts to assess the consequences of curricular alignment on student test performance. Finally, studies of curricular sensitivity in relation to dimensional structure of achievement tests are relevant to this study.

Curricular Alignment: Effects on Instruction

In 1975, the National Association of Elementary School Principals and others criticized the use of standardized achievement tests, stating, among other concerns, that these tests determine what is taught (Houts, 1975). In response, Rudman (1977) defended the use of standardized achievement tests and called for data to support the claims of critics. In recent years, however, such criticism has emanated from the ranks of measurement experts as well. As noted earlier, these criticisms can be grouped into four major categories:

narrowing the curriculum, deemphasizing complex thought, teaching test-taking skills rather than teaching content, and test score pollution, which reduces validity of achievement test scores. Each of these areas of concern is elaborated in the following sections of the review.

The Narrowing Curriculum

Koretz, Linn, Dunbar, and Shepard (1991) have reported that "in many states and districts across the country, testing programs have been transformed from a means of monitoring the progress of students into a mechanism for holding students, teachers, and school administrators accountable" (p. 1). They referred to this phenomenon as "high-stakes testing." In a high-stakes testing program, school-level personnel are often tempted to concentrate primarily on test items or objectives they know will appear on the test. According to Shepard (1989), teachers are pressured to see that test scores go up, whether or not a real increase in learning occurs. Consequently, other important aspects of the curriculum are ignored or lowered in priority in order to concentrate on tested objectives. Madaus (1983) stressed that this is especially true when minimum competency tests are introduced. He provided evidence from several research studies related to minimum competency tests to support the assertion that teachers will spend much of their time helping students to do well on a high-stakes test. According to Madaus, the impact on

curriculum of minimum competency tests in Ireland, Britain, and Florida have been studied. Outcomes of each research endeavor supported the contention that these programs lead to narrowing curriculum and instruction.

Frederiksen (1984) noted that almost all of the 50 states require some testing and that these requirements create a "feedback loop" in accountability. That is, testing requirements create pressure to teach tested objectives. Frederiksen conjectured that students prepare for measured objectives rather than experiencing a broad and rich curriculum.

Norm-referenced standardized achievement testing programs are usually considered to be "lower stakes" testing programs than state-mandated minimum competency tests which may be tied to promotion or graduation decisions. Nevertheless, due to the fact that achievement test results are now also being used for more far-reaching and important decision-making purposes than ever before, school-level personnel increasingly are tempted to devote instructional time to test-preparation activities and to alter the curriculum in order to cover objectives on these tests which will be tested. According to Airasian and Madaus (1983), standardized test results were once used primarily by teachers and were not terribly important for making major decisions. However, when accountability for the use of educational resources became a central concern, program

impact on student achievement test scores became the focal point. As more emphasis was placed on standardized achievement test results and as stakes (i.e., money, resources, career advancement) were tied directly to those results, school personnel became convinced that curricular alignment was both necessary and good. Even in a low-stakes achievement testing program, teachers and principals may be inclined to alter their curricula or to offer test preparation activities in order to help their students to "do their best" for a number of reasons (Darling-Hammond & Wise, 1985; Phillips, 1990; Shepard & Dougherty, 1991).

To investigate this phenomenon, Smith (1991) conducted a 15-month-long qualitative study of the effects of external tests on teachers in two schools within a district. She and her colleagues observed that teachers believe there is an emotional impact on students resulting from external testing. Smith suggested that teachers engage in a variety of activities designed to lessen the degree of emotional impact upon the students. She noted that "narrowing of curriculum happened before our eyes" (p. 10); namely, teachers began to leave out of their curriculum those objectives that were not to be tested.

Less Emphasis on Complex Thought in Instruction

Despite Ebel's (1965) demonstration that objective test items can be developed to test higher order thinking skills, there is a persistent view among educators that multiple

choice tests promote the teaching of skills at the lower level of the cognitive domain. Frederiksen (1984) contended that emphasis on easily measurable skills has resulted in decreasing performance on more complex cognitive skills in instruction. Voicing similar concerns, Haertel and Calfee (1983) reported that instructional activities in the classroom look more and more like test items. They found that publishers of educational materials provide worksheets that have the appearance of providing exercises for students that would help them to practice multiple choice test-taking skills. They concluded that educators "have begun to act as if every objective were amenable to testing, and to that particular form of objective testing that employs the multiple choice recognition format" (p. 123).

To determine the effect of test-based state education policies on classroom activities, Darling-Hammond and Wise (1985) interviewed 43 teachers from three large school districts. They reported that these teachers used fewer essay tests because they believed that these interfered with preparation for multiple choice tests. The researchers also noted that tested objectives were emphasized at the expense of untested objectives. Teachers also reported that they felt forced to concentrate on objectives that were most easily taught and tested. Often, this meant that in-depth discussion and analysis gave way to superficial

treatment and the breaking down of whole objectives into narrow subskills.

Smith (1991) described a situation in which third-grade teachers in a school were required to drill students in language skills due to the fact that their students' standardized achievement test scores had fallen short of "standard." To their language arts curriculum these teachers added multiple choice formatted drills, thus further reducing instructional time for "alternative topics or methods of teaching language" (p. 11).

According to Smith (1991) the following might be typical of the progression of the teaching of nontested subject matter over the course of a school year:

> As the school year progressed toward ITBS testing in April, [we saw a transition] from laboratory, hands-on instruction in science several days a week, to less frequent science out of textbooks . . . to no science instruction at all in the weeks before the test, to either no science at all or science for entertainment value during the ITBS recovery phase, to science instruction precisely tailored to the questions in the district criterion-referenced test, to no science at all. (p. 10)

Smith went on to note that similar patterns existed for writing instruction, whereas instruction in social studies and health was completely eliminated after December. Not only was the curriculum altered for these subjects, but "teachers tended to slight topics within math and reading that the test does not cover" (p. 10).

Shepard and Dougherty (1991) surveyed teachers in high-stakes districts in order to assess their perceptions of the effects of testing on their own classroom practices. Teachers of third-, fifth-, and sixth-grade students were randomly selected to receive the questionnaire. The total response rate was 42% (360 teachers). The results of a four-page questionnaire were first factor analyzed in order for the researchers to present responses for groups of similar items. Teachers reported that they felt pressure to see that their students' test scores improved and so they spent more classroom time emphasizing basic skills. They reported that less emphasis is given to nontested objectives. Although teachers identified some benefits of the use of standardized achievement tests, most believed that the negative effects outweigh the benefits.

An interesting contrast, however, is a study by Hall, Carroll, and Comer (1988). These researchers found that teachers who placed more value on the use of classroom and standardized test results tended to use more higher order questioning in their teaching. Reported correlations ($r = .25$ and $.26$) were low, but significant, at the level of .001.

Teaching Test-Taking Rather than Teaching Content

Teaching test-taking skills has a long history and has generally been regarded as effective. For example, Vernon (1954) presented evidence from several research studies

which indicated that providing practice materials based on test content serves to substantially enhance student test performance. Moore, Schutz, and Baker (1966) randomly assigned four ninth-grade classes to experimental and control groups in order to determine the effects of specific test-taking instruction given to the experimental group. The experimental group was instructed on methods of optimizing their scores based on the type of test (power test with no guessing penalty, speed test with guessing penalty, power test with guessing penalty, and speed test with no guessing penalty). They concluded that the students who had received test-taking instruction answered questions with the more appropriate strategies than did the control group. In yet another oft-cited study, Wahlstrom and Boersma (1968) demonstrated that "test-wiseness principles can be taught in the classroom and [can] lead to improvement in achievement test performance" (p. 419). They did so by randomly assigning ninth-grade students to experimental, control, and placebo groups, with the experimental group receiving instruction in test-taking skills. The placebo group served as an alternate control group to alleviate the problem of giving special treatment to the experimental group. The placebo group also received special treatment, but they did not receive instruction in test-wiseness. The experimental group outperformed the other two groups as expected.

In the 1960s and 1970s a number of authors produced books compiling hints and strategies for test-taking (e.g., Millman & Paulk, 1969), which increased the interest among educators in teaching test-taking skills and evaluating the effectiveness of the various strategies. Since then, several major reviews have been conducted to examine the effectiveness of coaching on standardized test performance. For example, Sarnacki (1979), in his review of relevant research, pointed to several researchers who have demonstrated "that training in various test-wiseness (TW) skills will improve test scores in subsequent standardized testing" (p. 265). He went on to point out that subject-matter was of no serious consequence in the teaching of test-wiseness skills. He found that "a large body of research has . . . demonstrated that many of the TW principles may be taught to examinees, with higher test scores resulting on a wide range of tests" (p. 274).

Bangert-Drowns, Kulik, and Kulik (1983) conducted a meta-analysis of the results of 30 studies on coaching for achievement tests. They found that "in 25 of the 30 studies in the sample, the coaching program had a positive effect on test performance, raising scores above the expected level" (p. 575). Amount of time spent coaching was positively correlated to effect size. Also, programs that included drills on content and especially those that used practice tests fared better than those that did not. However, amount

of training time had the most profound effect on effect size. The typical effect size was .2 standard deviations. Bangert-Drowns et al. concluded that coaching does help to boost achievement test scores, even though the effect size was typically small. However, in a subsequent meta-analysis, Kulik, Kulik, and Bangert-Drowns (1984) concluded that practice effects depend largely on how closely the practice tests parallel the actual tests. This factor appeared more important than the amount of practice time. This raises an interesting ethical issue about teaching test-taking skills to students (Mehrens & Kaminski, 1989).

Given this body of evidence, even in a low-to-moderate stakes testing situation, principals and teachers may perceive sufficient parental and community pressure to warrant the use of activities designed expressly to enhance students' test performances (Shepard & Dougherty, 1991). Madaus has voiced concern that "we don't really know how much time is spent on test preparation in a typical school year" (Brandt, 1989, p. 28). If substantial amounts of class time are spent on these activities at the expense of instruction in content areas, the result could be that children are learning less. Teachers felt that they spent an inordinate amount of time in test preparation activities, including the use of commercially produced test preparation materials. Several researchers (Haladyna et al., 1991; Smith, 1991) have suggested that as much as 100 hours of

instructional time is spent in testing, test preparation, and test recovery activities. This amounts to more than 3 weeks of school time in an academic year.

Test Score Pollution

In 1987 John Cannell (1987) published the results of a survey he had conducted in which he reported that all 50 states claim that their elementary-age students score above the national average on standardized achievement tests. This phenomenon became known as the "Lake Wobegon Effect" as psychometricians and measurement experts sought to explain and critique these findings and Cannell's methodology.

Phillips (1990) provided five possible reasons for the Lake Wobegon Effect. Three of the five explanations had to do with effects that would give the test user groups an advantage over the norm groups, namely, that (a) test users choose tests that most closely match their established curricula, (b) test users became more familiar with the tests over time, and (c) test user groups systematically exclude low achieving students from aggregate analyses. His other two explanations suggested that (a) "test users may be more motivated to do well than students in the norming sample, for whom the 'test doesn't count'" (p. 3) and (b) the self-selection process of the norming group may be more accurate reflections of current user groups rather than of national samples. Phillips (1990) went on to report that the U.S. Department of Education authorized a study to

replicate Cannell's 1987 survey and to address technical problems that were apparent in his report.

Linn, Graue, and Sanders (1990) reported the results of their surveys (both mail and telephone) intended to include a sample of school districts that would be nationally representative. These researchers asked state testing directors to provide reading and mathematics test results for every available grade for years 1985-86, 1986-87, and 1987-88. The 35 states that use achievement tests which allow comparisons with national norms reported that "the overall percent of students above the national median is greater than 50 in all of the elementary grades in both reading and mathematics for each of the 3 years studied" (p. 6). Linn et al. went on to emphasize that, although their findings were consistent with Cannell's (1987) earlier findings, the aggregate data mask the fact that many students score below the national average, and states and districts report these results as well.

Linn et al. (1990) also examined achievement test trends and addressed the issue of dated norms. State testing directors and some achievement test publishers indicate that student achievement performance is slowly but steadily increasing. Linn et al. suggested that the trend data that they examined could be interpreted as real achievement increases or as a reflection of "increasing familiarity with a given test form and more focused

instruction on the content of that specific form" (p. 11). They further noted that NAEP results indicate a trend of slightly increasing student achievement but not to the degree that has been demonstrated by other norm-referenced achievement tests.

Shepard (1990) interviewed state testing directors to ascertain their opinions about curricular narrowing and teaching to the test. Respondents to her survey indicated that they believe teachers place more emphasis on tested objectives simply because the objectives are tested. Shepard pointed out that curricular alignment to tests could indicate that students are learning important objectives. However, according to Shepard, the fact that the curriculum has been narrowed provides a source of test score pollution, as the national norm group did not have the advantage of the narrower, more focused curriculum.

Shepard (1990) also reported on some of the test preparation activities as described by survey respondents. "Most states do not . . . provide guidelines as to what constitutes appropriate test preparation" (p. 19). Activities ranging from the use of commercially developed practice test materials, to courses designed to cover objectives, to test preparation assemblies (pep rallies) were described.

Measurement experts recognize that achievement test scores are most useful for making inferences about the

domain of interest from which test items are samples (Mehrens & Kaminski, 1989). In order for generalizations to be legitimate, activities that would serve to prepare students for specific test items for the sole purpose of preparing the student for a specific standardized achievement test would alter the validity (and usefulness) of test scores. When Mehrens and Kaminski reviewed research studies using commercially published test preparation materials, they found that the use of these materials does not always produce substantial increases in students' scores. They concluded that some forms of measurement-driven instruction and test preparation activities are both "fruitless and fraudulent" (p. 21).

Perhaps the most significant contribution of Mehrens and Kaminski (1989) to this area was a list of types of activities that teachers may employ to raise student test scores and arrangement of these on a continuum ranging from "highly ethical" to "highly unethical." The types of activities identified included

1. General instruction on objectives not determined by looking at objectives measured on standardized tests;
2. Teaching test taking skills;
3. Instruction on objectives generated by a commercial organization where the objectives may have been determined by looking at objectives measured by a variety of standardized tests;
4. Instruction based on objectives that specifically match those on the standardized test to be administered;

5. Instruction on specifically matched objectives where the practice follows the same format as the test questions;
6. Practice on a published parallel form of the same test; and
7. Practice on the same test. (p. 16)

Haladyna, Nolen, and Haas (1991) specifically cited these activities as an important source of "test score pollution":

> (a) teaching test-taking skills, (b) promoting student motivation for the test, (c) developing a curriculum to match the test, (d) preparing teaching objectives to match the test, (e) presenting items similar to those presented on the test, (f) using commercial materials specifically designed to improve test performance, and (g) presenting before the test the actual items to be tested. (p. 4)

They placed these activities on a continuum from ethical to highly unethical. Activities (a) and (b) were classified as ethical, (c) through (f) were classified as unethical, and activity (g) was considered highly unethical. Haladyna et al. reported that "staggering" numbers of incidents of test score pollution were found in the reports they reviewed.

Nolen, Haladyna, and Haas (1992) surveyed teachers and administrators in the state of Arizona regarding test score use and test preparation activities. Approximately 41% (1,886) of the teachers and 47% (563) of the administrators responded to questionnaires that had been mailed to them. Most elementary school teachers and administrators indicated that test-taking skills are taught. Forty-one percent of the elementary teachers reported that they use commercially packaged test preparation materials. While 65% of the

teachers review tested topics prior to testing, 25% admitted to teaching vocabulary they know will be on the test, and, surprisingly, 11% and 10% use practice items from last year's test and the current test, respectively. Additionally, 8% of elementary teachers indicated that it is a common practice to ignore time allowances during testing. Also, 35% indicated that they review skills that will be covered on the next portion of the test.

Nolen et al. (1992) concluded that achievement tests are created as samples used to measure achievement of a larger domain. Test score pollution compromises that purpose by attempting to improve "performance on a small part of that domain, to inflate the estimate of achievement in the domain as a whole" (p. 14).

Empirical Studies of Curricular Alignment: Effects on Test Results

In 1986 Mehrens and Phillips examined two Michigan school districts that allowed school choice in curricula but varied in the emphasis placed on standardized test results. District-level supervisors rated school buildings on test-curriculum match. Textbook series used in the schools were also matched to standardized test content. After determining the degree of match between test content and curricula within the buildings, raw scores were statistically compared in order to determine differential impact of curricula on test results. The researchers concluded that differences in curricular offerings through

differential textbook series or curricular/test match have little impact on test results. The researchers admitted, however, that "it is possible that the lack of significant differences was a reflection of teachers stressing [test] content irrespective of the textbook" (p. 48).

Koretz, Linn, Dunbar, and Shepard (1991) examined the effects of a high-stakes testing program on achievement test results for a single district. They chose a large, high-poverty, urban district with a substantial black and Hispanic student population. Third-grade classrooms within schools were randomly sampled. Students were tested with the regularly scheduled, unmodified, commercial, norm-referenced achievement test that had been in use from 1986-1990. Later these same students were tested with the standardized achievement test that the district had used prior to 1986. Koretz et al. sought to determine the degree to which the standardized achievement test results were generalizable to an alternative achievement test measuring the same content. Due to the fact that the two tests' derived scores were based on different norm groups, adjustments were made to the results of the latter test. Koretz et al. concluded that "students in this district are prepared for high-stakes testing in ways that boost scores on that specific test substantially more than actual achievement in the domains that the test are intended to measure" (p. 2). They reached this conclusion by comparing

adjusted median grade equivalent scores on the mathematics subtests of the two tests that were administered. The researchers indicated that, by the spring of grade 3, student achievement, as measured by the standardized achievement test currently in use, is "overstat[ed] . . . by as much as 8 academic months" (p. 20). This interpretation, however, seems open to question because different tests may measure slightly different content. Furthermore, from a content validity perspective, it seems reasonable to expect that the test currently in use will better reflect the school's curriculum than a test that was formerly used.

Finally, Flexer (1991) administered an alternative test that had been designed for the study to students in the same district that Koretz et al. had studied. She sought to determine the extent to which students in this high-stakes district were being taught mathematics skills as opposed to being taught to do well on a standardized achievement test. The alternative test was designed to measure skills that were reported to have been taught before mid-March. Some of the skills tested on the alternative test were not tested on the district's standardized test, but were included in the district's stated curriculum objectives. This test also used a presentation of some items that were not in the multiple-choice format. Some of the items on the alternative test could be classified as an assessment of higher-order thinking skills. The researcher sought to

determine the extent to which teachers might be narrowing their curriculum by teaching to specific tested objectives as well as by teaching to the specific format of the standardized achievement test. After comparing the relative difficulties of the two tests by using an equating sample of students for whom both tests were novel, the researcher went about comparing student responses for the district in question. Flexer concluded that the achievement levels displayed by students in this high-stakes district "did not generalize to the alternative test" (p. 4). She also concluded that these students did not perform well on items that did not represent objectives that were tested on the district's standardized achievement test even though those objectives did appear in the district's written curriculum objectives guide.

Curricular Sensitivity and Dimensionality of Achievement Tests

If standardized achievement tests assess traits that are multidimensional in nature, it is possible that test preparation activities may alter test score validity without substantially altering test performance levels if patterns of student performance within the test are altered. An examination of total scores or item p-values alone cannot reveal this.

Measurement experts have tended to agree that most standardized achievement tests are probably multidimensional. Reckase (1979) contended that

standardized achievement tests are seldom unidimensional and are very often "factorially complex" (p. 208). He indicated that this is true due to the fact that standardized achievement tests are "not usually constructed using methodology designed to yield factor pure measures" (p. 208). Yen (1984) supported this point of view. Reckase (1985) has said that, often, test items measure a strong first factor and weaker subsequent factors. He suggested that mathematics word problems were the most obvious cases of this occurrence.

As an example of how "teaching to the test" may affect underlying dimensionality, consider a common item format for assessment of reading comprehension in which students are asked to infer from context the meaning of a word within a written passage. In the first year when the test is used, such items assess reading comprehension skill. In subsequent years, however, teachers may include these words on spelling and vocabulary lists, thus changing the underlying trait measured by these items and subtly altering the construct validity of the test. Traub (1983) also has provided an excellent example of how instructional differences may create additional dimensions underlying test score performance:

> Suppose we identify a subpopulation of examinees all having exactly the same ability in arithmetic computation. Now suppose this subpopulation is divided into two groups, one given intensive training on a subset of items referencing the ability (e.g., subtraction exercises), the other

> given intensive training on a different subset of items referencing the ability (e.g., addition exercises). If the effect of training is that subtraction items become relatively easier than addition items for the first group of examinees whereas addition items become relatively easier than subtraction items for the second group, then, after training, all pairs of items consisting of one addition and one subtraction item will be negatively correlated over this subpopulation of examinees. Prior to training, the correlations for these pairs would have been zero, . . . The consequences of training, according to this thought experiment, can be described as follows: Whereas before training only one number would have been sufficient to describe the relative abilities of the members of the subpopulation with respect to arithmetic computation, after training, two numbers for each of two abilities will be required. The two abilities will be addition and subtraction and the two numbers for each will differentiate the examinees who were trained on that ability from those who were not trained. (p. 60)

For some researchers, the most interesting thing about curricular differences and multidimensionality is the effect on the application of item response theory to estimating item and ability parameters. For others, the issue may be in terms of how different instruction affects the actual factor structure underlying the item responses to a particular test. Literature of both types is reviewed in the following sections.

Effects on IRT Parameter Estimation

Vojir and Shepard (1987) also believed that achievement tests are not designed as unidimensional measures. They investigated the effects of instructional variables on IRT item parameter estimates for an achievement test using a mathematics achievement test from the High School and Beyond

(HSB) Study. Students were classified into three groups based on self-reported mathematics instructional history. The groups contained (a) 4,043 students who had no college preparatory classes before the test, (b) 3,982 students who had only Algebra I before the test, and (c) 12,682 students who had at least one course beyond Algebra I before the test. These three groups constituted a pool from which subsamples were drawn.

The first sample of 3,300 students was drawn in order to determine whether instructional history is related to item parameter estimates. The subsample contained approximately 1,100 students from each of the three aforementioned instructional history groups. The second sample of 2,200 students was drawn from the pool of students who had taken courses past Algebra I before the test. This subsample was used to provide a baseline for item bias analyses. The third sample of 2,200 students was drawn from the pool of students who had courses past Algebra I. The purpose of this subsample was to determine whether group test-score distributions influence bias index outcomes. The fourth sample of 3,300 students was drawn to represent a simulated change in high school graduation requirements requiring Algebra I. The researchers hoped to discover what effect multidimensionality would have on achievement test results given a simulated change.

Subsample groups from samples 1, 2, and 3 were analyzed using three-parameter IRT models. The c's were estimated for the whole sample and then remained fixed for subsequent analyses. Item characteristic curves (ICC) were estimated and nine bias indexes were used to compare subsample groups. Among the nine indexes were (a) Lord's asymptotic chi-square; (b) estimated unsigned area between ICCs; (c) estimated signed area between ICCs; (d) two weighted sum-of-squared differences; (e) two unweighted sum-of-squared differences; and (f) simple differences between estimated parameters.

The second sample was used to estimate what one might expect of bias indexes from random data. If at least four of the bias indexes were larger than expected (based on their baseline values), the items were judged to perform differently than the other 20 test items. For the comparison of the no-college-preparatory-courses group and Algebra-I-only group from Sample 1, ICC differences were found for two items. For the comparison of the no-college-preparatory-courses group and the past-Algebra group, ICC differences were found for 11 items. For the comparison of the Algebra-I-only group and the past-Algebra group, ICC differences were found for 10 items. When the researchers compared instructional groups from Sample 3, ICC differences were found for 10 items. From these results, the researchers concluded that the achievement test was

multidimensional and that one factor affecting item characteristic curves is instructional history differences.

Cook, Eignor, and Taft (1988) studied the effects of recency of instruction on student performance on biology achievement tests. Students who had taken a biology achievement test in the fall or spring were used as subjects for this study. Students who took the test in the spring, generally had more recent biology instruction than had students who were tested in the fall. Data for this study came from three administrations of biology achievement tests (two forms) in both spring and fall. Random samples of 2,500 to 4,000 students who took tests of 99 items (old form) or 95 items (new form) were used. The two test forms had 58 items common to both. Comparisons were made for the 58 common items on test form-sample combinations (i.e., "spring new form"--"fall old form" and "fall new form"--"fall old form").

From the 58 common items, three other subtests were developed. One subtest was formed from the 29 items that were thought to be least stable because they changed most when the fall old form administration results were compared to the spring new form results based on equated delta plots. The second subtest was formed from the 29 items that were thought to be most stable on the fall-spring delta plots comparisons. The third subtest of 36 items was formed based on the opinions of measurement experts who deemed those

items least likely to change based on content considerations. The additional subtests were analyzed using delta equating of the "spring new form"--"fall old form" and "fall new form"--"fall old form" comparisons.

Item response theory item parameters estimates were made using a 3-parameter model for each of the three total tests. Item characteristic curves were transformed to a common scale using the four aforementioned common sets of items.

Cook et al. examined the stability of the delta plots by correlating the new form and old form deltas for each of the four subtests of common items. They found that correlations for "fall old form"--"spring new form" comparisons ranged from .74 to .92. On the other hand, correlation coefficients for "fall old form"--"fall new form" comparisons ranged from .98 to .99 on the four subtests of common items.

These researchers also examined the stability of the IRT item difficulty parameters for all form-subtest combinations. They correlated the three item parameter estimates, producing coefficients for "fall old"--"spring new" comparisons ranging from .52 for the c parameter to .93 for the b parameter. Again the "fall old"--"fall new" comparisons produced higher correlations ranging from .62 for the c parameter to .98 for the b parameter. Item difficulty estimates were more stable when the two fall

samples were compared than when the spring sample was compared to the fall sample.

Cook et al. concluded that the common items measured different "attributes" when spring and fall samples were compared due to the fact that these populations differed substantially in OTL. When fall samples were compared, the tests appeared to measure similar things. Item difficulty parameters (using both conventional and IRT methods of estimation) differed for student groups tested in the fall when compared to those tested in the spring.

Item parameter invariance means that, excepting sampling error, item characteristic curves should be the same across curriculum groups. These researchers indicated that if item characteristic curves differ across curriculum groups, this implies that the assumption of unidimensionality does not hold. Miller and Linn (1988) also studied the effects of differing instruction on item characteristic curves, using 6,788 eighth grade U.S. students' arithmetic and algebra test data from the Second International Mathematics Study (SIMS). A student's opportunity to learn (OTL) was defined in terms of whether or not the item content was covered within the previous academic year.

Students were classified into groups based on teacher-reported OTL data. The researchers performed cluster analyses based on the OTL data for all of the items.

Students were assigned to the cluster into which their teachers had been assigned. Each student took a common core of 40 items and one rotated-form of the test (35 items). There were a total of four rotated-forms of the test. For each curriculum cluster, item parameters were estimated separately using samples of approximately 1,000 students. The "a" and "b" parameters were equated across groups using a linear equating.

Miller and Linn used two indexes to describe item invariance. The first was the unsigned sum of squared differences (USOS) for an item's characteristic curve when two curriculum clusters were compared. The second index was the signed sum of squared differences (SSOS) for an item's characteristic curve when two curriculum clusters were compared. The SSOS differed from the USOS in that the SSOS values took into account the direction of the bias using the sign of the value.

Baseline data for comparison purposes were established by computing USOS and SSOS from two samples drawn from the same curriculum cluster. By comparing SSOS and USOS statistics to a cut-off of the baseline mean plus two standard deviations, they found that 24 of the 62 arithmetic items would have been deemed invariant in one comparison and 18 items in a second comparison. Of the 31 algebra items at least 6, 11, and 5 items on the respective subtests would

have shown item parameter invariance using either USOS or SSOS.

Miller and Linn concluded that differences in OTL led to a lack of invariance in item characteristic curves. They noted, however, that "OTL or any single measure of instructional coverage or practice, probably only scratches the surface of differences in how, why, and in what order a topic is taught in the class" (p. 216).

Muthen (1988) contended that the use of IRT modeling may be suitable for relating responses on an achievement test to opportunity to learn. He used the Second International Math Study (SIMS) U.S. eighth-grade data from fall and spring testings for his analysis (approximately 4,000 students who answered 40 items each). Muthen categorized OTL based on two questions asked of the teachers for each item. OTL was defined in terms of three levels: (a) no OTL, (b) prior OTL, and (c) this year OTL. He used an IRT model that allowed for OTL variation. This means that the researcher would not falsely detect "item bias" when, in fact, group differences were attributable to OTL. Subsequently, Muthen, Kao, and Burstein (1991) used LISCOMP (Muthen, 1987) in a confirmatory factor analysis model in order to take into account that item difficulty estimates vary as a result of student OTL levels, multidimensionality, and student achievement levels. They noted differing effects of OTL on achievement levels due to the fact that

some items may be more sensitive to exposure than other items. In other words, some items may be "definitional" in nature, such that exposure to the correct definition makes the item easy for most students. On the other hand, some items may be so difficult that mere exposure is not enough to allow for a large percentage of examinees to get the item correct. They concluded that much more work needs to be done in the area of investigating the instructional sensitivity of items on standardized achievement tests and the relationship of that sensitivity to IRT models. It is important to note here that, while Muthen's procedures hold promise for future work in this area, the relatively large subject-to-item ratio required to achieve stable results with this procedure makes it impractical for use with sample sizes available in single districts with small or moderate student enrollments. The number of students available at any one grade level would be insufficient and typical test length too great to produce stable results.

Effects on Factor Structure

Phillips and Mehrens (1987) used the test scores from their 1986 study with linear factor analysis in order to determine if curricular differences within a district affect the dimensionality of achievement test scores. They applied linear factor analysis of inter-item tetrachoric correlation matrices obtained from examinee item responses. The researchers evaluated the results by examining relative

differences in percentages of variance on the first factor and subsequent factors that were significant. They used values for the first five factors for each subtest by ratings or textbook series groups. Phillips and Mehrens indicated that they observed a "clearly dominant factor in some cases" (p. 8). They did not detect large differences in first factor magnitudes across ratings groups nor across textbook series groups for each district. They also used IRT item bias analyses in order to detect item performance differences among students who receive different instruction before taking the same standardized achievement test. The results of this research indicated that, for schools that adhere closely to achievement test objectives compared to those that do not, there is little difference in dimensionality underlying item performance. This was also true for schools using different texts but the same standardized test.

Phillips and Mehrens (1988) again used data from the same districts to examine the effects of curricular differences on achievement test data at the item and objective levels. They used classical item p-values and delta plots to examine the differential effects of curricular/test match on test scores for each objective. They also examined the item p-values and delta plots to examine the differential effects of curricular/test match on test scores for each objective. Finally, they examined the

item p-values to determine if curricular differences affected those values. The authors used linear factor analysis of interitem tetrachoric correlation matrices to examine differences in factor loadings for different curricula. The results of this research were similar to their earlier studies in that they found that curricular differences had little impact on standardized achievement test performance at the item and objective levels.

Caution seems appropriate, however, in accepting the findings of Mehrens and Phillips uncritically. Limitations of their study include the failure to document how long the test had been in use in these districts prior to the study, determination of match between test and curriculum by having district office personnel rate the test as a whole rather than examining specific items by classroom teachers who actually control the instruction, and failure to apply factor analytic procedures that would permit testing the goodness-of-fit of the factor structure across different curricular groups.

Summary

A number of researchers have expressed concerns and offered substantial evidence to indicate that teachers alter their curriculum to improve student test performance (Darling-Hammond & Wise, 1985; Madaus, 1983; Phillips, 1990; Shepard & Dougherty, 1991). There are, however, mixed findings to indicate that these teaching practices affect

student test performance (Flexer, 1991; Koretz et al., 1991; Mehrens & Kaminski, 1989; Wahlstrom & Boersma, 1968).

A growing body of literature also shows that test dimensionality is affected by curricular differences on a nationwide level. Typically, these studies focus on how opportunity to learn (OTL) has affected item parameter and ability estimates obtained in item response theory.

Muthen and several of his colleagues (Muthen, 1988, 1989; Muthen, Kao, & Burstein, 1991) have used LISCOMP (Muthen, 1987) to examine the effects of previous opportunity to learn (OTL) on test item performance. This research method allowed for an OTL factor for each test item. Its focus was on detection of item bias resulting from differential instruction. However, these researchers ignored other aspects of test score pollution such as motivational factors, practice test effects, and teaching test-taking skills (Haladyna, Nolen, & Haas, 1991; Mehrens & Kaminski, 1989). The present study has included some of these other factors in a broader view of "teaching to tests." Muthen also used extremely large data sets with large examinee-to-item ratios. It is possible that smaller data sets such as those which would typically be available from middle-sized public school systems would result in more unstable estimates if LISCOMP were to be used.

Mehrens' and Phillips' studies (Mehrens & Phillips, 1986, 1987; Phillips & Mehrens, 1987) seem promising in

approaching a broader definition of measurement-driven instruction and the impact on test results. These researchers documented curricular differences by noting differential textbook use and by ratings regarding closeness of test-curriculum match. They concluded that differential textbook use had little impact on test results. The results of their factor analyses led to the conclusion first factor magnitude differences across groups were not substantial. In the present study the researcher used a population similar to one of the populations studied by Mehrens and Phillips. Mehrens and Phillips obtained data from a middle-sized school district with a student population of about 23,000 (1600-1700 per grade) with approximately 40% minority students. They focused on grades 3 and 6. In the present research test item response data and teacher questionnaires were obtained from a middle-sized school district with a student population of about 25,000 (38% minority), focusing on grades 2 through 5. The present research was designed to overcome the major limitation noted for Mehrens and Phillips (1986, 1987) with regard to the number of years the standardized test was in use before the study began by utilizing data from the first year of implementation of a new testing program. Classroom teachers were asked to report on test preparation activities that occurred for their students in each of the 3 focal years of the study.

CHAPTER 3
METHODOLOGY

The purpose of this study was to examine the effects of test preparation and curricular alignment activities over a 3-year period in a middle-sized school district in north-central Florida. The study focused on the first year of the use of a standardized achievement test through the third year of its use. The effects of curricular alignment and test preparation activities were examined using multiple regression, exploratory factor analysis, and confirmatory factor analysis. Overall test performance and factor structures of student responses were compared for the same teachers in years when teachers did and did not expect their students to be tested.

Study Setting

Students and teachers from a middle-sized school district in north-central Florida provided data for this study. This school district had a unique standardized testing experience during the focal years of this study that rendered it especially suitable for this investigation. The California Achievement Test (1985 ed., Form E) was adopted for first use during the spring of 1986-87 school year. The same test was then used during the 1987-88 school year. In

October, 1988, the school board members voted to deemphasize standardized test results, announcing that children in kindergarten, third, fourth, sixth, and eighth grades were not to be tested with the California Achievement Test (CAT) except when necessary to meet federal and state compensatory education requirements. However, in the spring of 1989, just a few days prior to testing, this same school board voted to reinstate testing at all grade levels except kindergarten.

The feature of particular interest for this study was that teachers of third and fourth grade who believed that their students would not be tested with the CAT were surprised just a few days before testing to learn that, in fact, their students would be tested. This offered the opportunity to contrast performance of these teachers' students in the first year of testing, the second year of testing when the test was expected, and the third year of testing when testing was not expected. Table 1 provides an overview of this situation.

Sample

For this study the researcher obtained data for elementary school classes of a middle-sized metropolitan school district in north-central Florida during the 1986-87, 1987-88, and 1988-89 school years. Classes of teachers who taught the same grade level from 1986-87 through 1988-89 constituted the sample of interest. Approximately 24

teachers per grade level were selected. Each class had approximately 25 students for a total of 600 students per year per grade and 1800 students per grade over the 3-year period (see Table 2).

Table 2

Number of Teachers and Students in Sample by Year and Grade

Grade:	2	3	4	5	Total
Teachers:	27	24	22	21	94
1987 students	632	534	605	596	2367
1988 students	650	574	633	618	2475
1989 students	618	573	596	589	2376
Total students	1900	1681	1834	1803	

Instruments

Principal interviews and a teacher questionnaire were developed by the researcher. Both instruments were reviewed by district-level curriculum staff members, two university measurement experts, the district research director, two principals, and the district school superintendent. Each of the instruments was piloted and revised before being used district-wide.

All elementary school principals in the district were interviewed by the researcher during the late summer and

fall of 1989. The purpose of these interviews was to determine the types of test preparation activities that were being implemented on a school-wide basis or at particular grade levels. A teacher questionnaire was developed based on the results of these interviews and a review of literature. On this questionnaire, teachers indicated whether specific test preparation activities had occurred for their students for each of the 3 focal years in question. The activities fell into three broad categories: "Motivational/Psychological," "Curricular Alignment," and "Test-taking skills." The Motivational/Psychological category included such specific activities as "test busters" assemblies, relaxation exercises, and motivational speeches during morning announcement time. Curricular alignment included the reordering of the teaching of specific text chapters in order to introduce or to emphasize certain skills prior to administration of the standardized achievement test. Test-taking skills activities included teaching students proper bubbling techniques and methods whereby foils could be eliminated when a student was unsure of the correct answer.

The California Achievement Test (1985 ed., Form E) was administered in grades 1-5. All students took the Total Reading, Language Expression, and Total Mathematics subtests. The norm group for the CAT consisted of 300,000 K-12 students in the fall, 1984, and spring, 1985. The

reliability estimates, reported in terms of KR-20, for grades 2 and above range from .80 to .90.

Design and Data Collection

The aforementioned principal interviews were standardized in that all principals were asked the same questions. The researcher was able to talk candidly with elementary building principals regarding past test preparation activities. Principals were informed that the researcher would be allowed to keep all of their responses confidential and that individual schools would not be identified when data were analyzed and reported.

Although teachers' names were printed on the teacher surveys that were sent via district courier to teachers at their schools, they received the same assurance of confidentiality as principals had received. This teacher questionnaire was sent out during the late winter, 1989, and returned to the researcher by early spring, 1990. A cover letter from the district's research director indicating that the school board and the superintendent were interested in the results of this project helped to elicit a 100% response rate that allowed the inclusion of classes of all teachers who were identified as part of the sample.

Item response data for the 1987, 1988, and 1989 CAT test administration were available on computerized data bases. The researcher used computer program merging capabilities to develop a data base that included only the

item responses of students in the classes of interest. To this data base were added the results of the principal and teacher surveys. Subsequently, the data base included for each class in each year each student's item responses, the grade level, and pertinent demographic information on each student. This allowed the researcher to aggregate data at the student, class, school, grade, and district levels.

Method of Analysis

Data were analyzed by grade level to determine if differences in amount of test preparation activities existed for the years and grade levels dependent on whether or not teachers expected their students to be tested. It was expected that teachers who did not expect their students to be tested during spring, 1989, would have participated in fewer test preparation activities during the 1988-89 school year, and that this would have subsequent impact on both overall performance levels and factor structures.

In order to answer question 1, a two-way analysis of variance was used. It was expected that there would be a grade-level by year interaction with regard to the extent to which test preparation activities took place. Specifically, second and fifth grades were expected to participate in similar amounts of test preparation activities over the 3-year period, whereas third and fourth grades participated in fewer test preparation activities between years 2 and 3 than they did from year 1 to year 2.

Multiple regression techniques were employed to determine the effects of differential test preparation activities on standardized achievement test scores. In addressing the second research question, class means served as the dependent variable. Independent variables were prior level of achievement, participation/no participation in test preparation activities for each of the years in question. The model used was as follows:

$$Y'_a = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5$$

Y'_a = predicted class mean for year a,

X_1 = prior achievement,

X_2 = curricular alignment activities score,

X_3 = motivational activities score,

X_4 = test-taking skills score, and

X_5 = practice in subject areas score.

Beta weights were examined to determine if any of the four test preparation activities (X_2, X_3, X_4, X_5) significantly affected class means. This gives the relative importance of the activity, as scaled, in the prediction of achievement.

In addressing questions 3 and 4, a repeated measures design was used. Class means for years 1, 2, and 3 were

compared. It was expected that for question 3 (second and fifth grades), student test performance from year 1 to year 3 would improve. For question 4 (third and fourth grades), the expectation was that student test performance would improve from year 1 to year 2 and would decrease from year 2 to year 3.

For questions 5 and 6, an inter-item tetrachoric correlation matrix was estimated for each year. This was analyzed using an exploratory, common-factor method. Numbers of factors were determined based on a scree procedure. Items were assigned to factors when they had loading values of .30 or greater. Resulting factor structures were compared based on whether the same number of factors were identified from year to year and similarity of item assignment to factors. For grades 2 and 5, respectively, factor loading patterns were expected to vary somewhat more than for grades 3 and 4 when year 1 was compared to years 2 and 3.

The steps in the application of confirmatory factor analysis for comparisons of years were as follows:

1. Identification of a meaningful baseline model. In this case, for each grade level, the baseline factor structure was obtained from an exploratory factor analysis of the responses of students at that grade during the first year the test was administered.

2. This structure was considered invariant for the subsequent years by constraining the pattern of fixed and free elements of Λ_x to fit that of the first model and comparing the goodness-of-fit of the data from the subsequent year to the baseline model.

CHAPTER 4
RESULTS

Research Questions

The following research questions were addressed in this study:

1. How did test preparation activities change over a 3-year period following the adoption of a new achievement test? Did the amount of activities differ by grade levels dependent on whether there was an expectation to test in the third year?

2. Was overall student test performance in mathematics affected by teacher curricular alignment or test preparation activities?

3. At second and fifth grades, where teachers expected to be tested each year, were there increases in student math scores of the 3-year period?

4. At third and fourth grades, was there a decrease in student math scores in the year when teachers did not expect to be tested?

5. For each grade level, did the factor structure of dichotomous item responses on the mathematics test remain stable over the 3-year period (in terms of number of

factors, similarity in patterns of item loadings, and magnitude of goodness-of-fit indexes)?

6. For each grade level, was the factor structure of the skill-area subtest scores stable over the 3-year period (in terms of number of factors, similarity of subtest loadings, and goodness-of-fit indexes)?

Question 1: Changes in Test Preparation Activities Over Time

A questionnaire was sent to each teacher in the school district who had been targeted for the sample. The criterion for inclusion in the sample was based on the fact that he/she had taught the same grade level during the 3 focal years of the study. The 94 teachers in the sample were asked to respond to questions related to curricular alignment and test preparation activities that took place for their students during the 3 years. Questions dealt with activities such as curricular alignment, test-taking skills, and motivational activities. Within each category of test preparation activities, there was a possibility of six specific activities. Thus, each teacher received a score ranging from 0 to 6 for each of three categories of each of the activities during each of the 3 years in question. Additionally, teachers indicated whether or not students had participated in subject area practice activities for each year. Teachers received a score of 0 if students did not participate in a subject area practice activity and a score of 1 if students did participate in a subject area practice

activity. Presented in Table 3 are average scores for each year and grade level for each type of activity. These averages seem to indicate that more test preparation activities occurred in year 2 than in years 1 or 3. This is consistent with the expectation that teachers engaged in most test preparation activity in the year immediately following their first exposure to the test.

A two-way analysis of variance using Huynh-Feldt's corrected degrees of freedom indicated a significant year-by-grade interaction, $F(6,176) = 27.25$, $p<.0001$, on test preparation activities. Subsequent pairwise comparisons of grade levels for each year indicated there were no significant differences in amounts of test preparation activities among grades 2 through 5 in 1987, the first year of the program when the test was unfamiliar to all. During 1988, there were no significant differences in amounts of test preparation activities among grades 2 through 5 when all teachers had seen the test in the previous year and knew that their students would be tested. However, in 1989, there were some significant differences. In 1989 teachers in grades 2 and 5 expected their students to be tested, but teachers in grades 3 and 4 were not expecting their students to be tested. Tukey's studentized range (HSD) test, controlling the experimentwise error rate, indicated the following significant differences.

Table 3

Average Number of Test Preparation Activities by Grade and Year

Grade	Year: Activity	1987 Mean	(SD)	1988 Mean	(SD)	1989 Mean	(SD)	Max
2	Motivational	.5	(.7)	3.0	(1.7)	2.3	(1.2)	6
	Test-taking skills	1.4	(1.2)	3.3	(1.2)	2.9	(.9)	6
	Curricular alignment	.3	(.4)	3.6	(1.5)	3.6	(1.3)	6
	Subject area practice	.2	(.4)	.7	(.4)	.6	(.4)	1
3	Motivational	.9	(1.3)	3.4	(1.6)	.3	(.7)	6
	Test-taking skills	1.5	(1.2)	3.8	(1.1)	.2	(.3)	6
	Curricular alignment	.3	(.4)	3.9	(1.7)	.7	(.7)	6
	Subject area practice	.5	(.5)	.7	(.4)	.4	(.5)	1
4	Motivational	.9	(1.0)	3.6	(1.8)	.1	(.2)	6
	Test-taking skills	1.7	(1.7)	3.3	(1.4)	.1	(.3)	6
	Curricular alignment	.4	(.5)	3.9	(1.7)	.3	(.6)	6
	Subject area practice	.3	(.4)	.7	(.4)	.4	(.5)	1
5	Motivational	1.2	(1.3)	3.9	(1.7)	2.4	(1.4)	6
	Test-taking skills	1.8	(1.3)	3.2	(1.2)	3.1	(1.2)	6
	Curricular alignment	.6	(.5)	3.7	(1.4)	2.8	(1.2)	6
	Subject area practice	.5	(.5)	.8	(.4)	.8	(.4)	1

1. Grade 2 differed significantly in amount of test preparation activities from grades 3 and 4 ($p<.05$).

2. Grade 5 differed significantly in amount of test preparation activities from grades 3 and 4 ($p<.05$).

3. Grades 3 and 4 did not experience significant differences in amounts of test preparation activities.

4. Grades 2 and 5 did not experience significant differences in amount of test preparation activities.

Thus, in year 3, beliefs about whether one's students would take the achievement test led to differences in classroom behavior involving activities related to test preparation and curricular alignment (see Table 4).

Table 4

Significant Differences in Test Preparation Activities during Year 3 between Pairs of Grade Levels

Grade	Compared to	Grade	
2	-	3	*
2	-	4	*
2	-	5	
3	-	4	
3	-	5	*
4	-	5	*

*indicates that the difference in levels of activity is significant ($p<.05$).

Question 2: Effects of Curricular Alignment, Test Preparation Activities on Mathematics Test Scores

Students in grades 2 through 5 took the mathematics portion of the California Achievement Test during the spring of the 3 focal years of the study. For each of those years students were prepared for test taking in a variety of ways that have been described previously. In order to determine

whether any of these types of test preparation activities significantly affected student test scores, multiple regression techniques were employed. Class means served as the dependent variable. Independent variables were prior level of achievement (pretest mean), score on the motivational activities portion of the teacher questionnaire, score on the test-taking skills preparation portion of the teacher questionnaire, score on the curricular alignment portion of the teacher questionnaire, and participation/no participation in subject area practice activities. The following model was used:

$$Y'_a = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5$$

Y'_a = predicted class mean for year a,

X_1 = pretest mean,

X_2 = motivational activities score,

X_3 = test-taking skills score,

X_4 = curricular alignment score,

X_5 = participation/no participation in subject area.

For the first year of test administration the only activity that had a significant relationship to achievement test scores was curricular alignment activities at grades 5, $t = 3.4$, $p<.05$. The only curricular alignment activity that was used during the first year of test administration was that some teachers reported that they had used results of the previous year's test to determine objectives that should be

emphasized during the current year. Apparently, the teachers believed that objectives on which students showed weaknesses should be emphasized, regardless of the fact that a new standardized test would be used. This seems to reflect a view that they felt the standardized test results identified legitimate instructional weaknesses that should be remedied.

For the second year of test administration, pretest scores proved to be significantly related to mathematics achievement for grades 3 through 5. Additionally, test-taking skills activities were significantly related to mathematics achievement at grade 2, $t = 2.328$, $p<.05$. Motivational activities were significantly related to mathematics achievement for students in grade 4 during the second year, $t = 2.227$, $p<.05$.

Pretest scores were significantly related to mathematics achievement at all grade levels in year 3 as well. During year 3, curricular alignment activities were significantly related to mathematics achievement at grade 2, $t = 2.310$, $p<.05$. At grade 4, during year 3, motivational activities were significantly related to mathematics achievement, $t = 2.370$, $p<.05$.

Presented in Tables 5 to 8 are data relevant to statistical tests of significance for comparisons of mean raw scores for each grade and year. R-squared and F statistics are reported for the full regression model and, corresponding

t-values for each variable included in the model are also reported.

Table 5

Grade 2 Comparison of Mean Raw Scores by Year

	Model			t for Variables				
Year	R^2	F	N	Pretest	MA	TTS	CA	SAP
1987	.18	1.23	27	----	1.26	.61	- .59	- .56
1988	.41	2.87*		1.39	-1.14	2.32*	- .18	1.07
1989	.65	7.92*		4.48*	-1.41	1.44	2.31*	-1.21

*p<.05 (MA = motivational activities; TTS = test-taking skills activities; CA = curricular alignment activities; SAP = subject-area practice activities).

Table 6

Grade 3 Comparison of Mean Raw Scores by Year

	Model			t for Variables				
Year	R^2	F	N	Pretest	MA	TTS	CA	SAP
1987	.29	1.96	24	---	.72	.37	1.50	.26
1988	.74	10.47*		6.52*	- .55	.28	.03	.08
1989	.51	3.84*		3.58*	-1.07	.53	.69	-.62

*p<.05 (MA = motivational activities; TTS = test-taking skills activities; CA = curricular alignment activities; SAP = subject-area practice activities).

Table 7

Grade 4 Comparison of Mean Raw Scores by Year

	Model			t for Variables				
Year	R^2	F	N	Pretest	MA	TTS	CA	SAP
1987	.22	1.21	22	----	.23	.08	1.60	.009
1988	.74	9.31*		5.13*	2.27*	-1.91	-1.09	1.05
1989	.65	6.19*		2.50*	2.37*	-1.07	.37	1.65

*p<.05 (MA = motivational activities; TTS = test-taking skills activities; CA = curricular alignment activities; SAP = subject-area practice activities).

Table 8

Grade 5 Comparison of Mean Raw Scores by Year

	Model			t for Variables				
Year	R^2	F	N	Pretest	MA	TTS	CA	SAP
1987	.49	3.94*	21	----	-.02	.14	3.43*	-.66
1988	.85	17.37*		8.32*	.87	.10	- .63	-.25
1989	.88	22.33*		8.20*	-.06	1.55	1.55	-.40

*p<.05 (MA = motivational activities; TTS = test-taking skills activities; CA = curricular alignment activities; SAP = subject-area practice activities).

Questions 3 and 4: Changes in Mathematics Subtest Scores over a 3-Year Period

Presented in Table 9 are mean raw scores and standard deviations for classes by grade and year on the Mathematics subtest of the California Achievement Test. As expected,

Table 9

Mean Raw Score (Standard Deviation) for Each Grade and Year on Mathematics Subtest, CAT

Year:	1987		1988		1989	
Grade	Mean	SD	Mean	SD	Mean	SD
2	42	6	50	3	49	3
3	55	6	72	7	71	5
4	74	9	80	9	78	6
5	77	7	80	7	77	9

raw scores increased, on the average, the second year the test was administered. The expectation for the third year of testing was that scores for grades 2 and 5 would remain the same, on the average, due to the fact that no change in the testing schedule had been announced for those grade levels. Average raw scores for grades 3 and 4, however, were expected to drop due to the fact that for most of the school year, these teachers believed that their students would not be tested. Repeated measures analysis techniques were employed to determine the significance of the average raw score

changes reported in Table 5. Follow-up analyses for pairwise comparisons employed Bonferroni t-test analyses controlling family-wise error rates.

At grade 2, the raw score changes for years 1, 2, and 3 were significant, $F(2,52) = 35.08$, $p<.05$. Follow-up analyses indicated that year 1 average raw scores differ from year 2 significantly, $t(26) = 7.71$, $p<.05$. Year 2 average raw scores were not significantly different from year 3, but year 1 class means differed significantly from year 3, $t(26) = 5.52$, $p<.05$.

At grade 3, the raw score changes for years 1, 2, and 3 were significant, $F(2,44) = 102.66$, $p<.05$. Follow-up analyses indicated that year 1 average raw scores differed significantly from those of year 2, $t(22) = 14.75$, $p<.05$. Raw score averages between years 2 and 3 did not differ significantly. The difference between raw score average for years 1 and 3 was significant, $t(22) = 12.08$, $p<.05$.

At grade 4, the changes in raw score averages for years 1, 2, and 3 were significant, $F(2,42) = 6.56$, $p<.05$. When comparing raw score averages for years 1 and 2, significant differences were found, $t(21) = 5.10$, $p<.05$. Raw score averages for years 1 and 3 did not differ significantly. At grade 5 the raw score changes for years 1, 2, and 3 did not differ significantly.

Provided in Figure 1 are visual representations of the raw score means for all grade levels. Expectations for

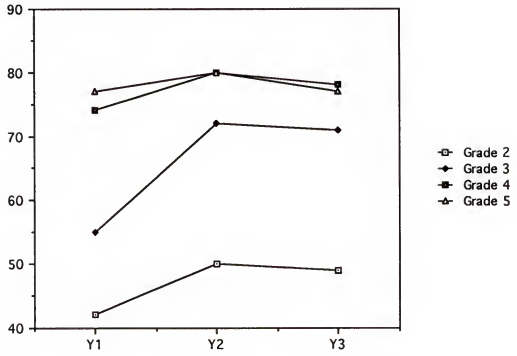


Figure 1. Raw score means on CAT mathematics test of three grade levels over 3 years.

grade 5 were not met, as there were no significant differences in average raw scores for the 3 years. Expectations for grade 2 were met, in that year 1 average raw scores differed significantly from those in year 2. But year 3 average raw scores (though lower than year 2) did not differ significantly from year 2 and were significantly different from those in year 1. Although grade 3 classes experienced the greatest increase from year 1 to year 2, the drop in average raw scores from year 2 to year 3 was not significant. Grade 4 average raw scores increased significantly from year 1 to year 2 as was expected. However, in year 3, the average raw score was not significantly lower than in year 2 nor was it significantly higher than year 1.

Questions 5 and 6: Factor Structures of Items Responses over a 3-Year Period

An inter-item tetrachoric matrix was estimated for student responses to items on the Mathematics Computation and Mathematics Concepts and Applications subtests for each year and each grade level. These matrices were analyzed using a purely exploratory common factor method. The number of factors was determined based on a scree procedure. Presented in Tables 10 and 11 are the numbers of factors for each grade level by year as determined by examination of scree plots of eigenvalues and of eigenvalue differences. (See Appendix B.)

Table 10

Math Computation: Numbers of Factors by Grade and Year

Grade	1987	1988	1989
2	3	4	3
3	4	5	4
4	2	4	3
5	3	4	3

Table 11

Math Concepts and Applications: Numbers of Factors by Grade and Year

Grade	1987	1988	1989
2	3	5	4
3	4	4	4
4	3	3	3
5	3	3	3

Factor structures over the 3-year period appeared to be somewhat unstable, but there was a trend for an increase in the number of factors from the first to the second year. In Math Computation, the number of factors dropped back to the

original number in the third year. In Math Concepts and Applications, the increase in number of factors was observed for grade 2, but the number of factors for grades 3, 4, and 5 remained stable over all years.

After determining the number of factors for each grade level in the base year (year 1), factors were rotated using the varimax technique, and item factor loadings were estimated. Items were assigned to factors when they had factor loadings of .30 or greater.

Grade 2 Results

Exploratory factor analysis techniques on second-grade test items yielded results that indicated that the Mathematics Computation subtest could be explained by three major factors in the base year. An examination of the items which loaded on each factor led to the conclusion that the factors are (a) addition without carrying, (b) subtraction with no regrouping, and (c) addition/subtraction with carrying/regrouping. An examination of scree plots and eigenvalue differences indicated that in year 2 four factors could best explain the Mathematics Computation portion of the test. An examination of item factor loadings suggested that the factors were (a) addition/subtraction with no carrying/regrouping, (b) subtractions with no regrouping, (c) addition/subtraction with carrying/regrouping, and factor 4 had only one item in horizontal format. This item had a loading of .84 on factor 4 while it loaded on factors 1

through 3 with values of .2, .3, and .1, respectively. The item was the only subtraction item requiring regrouping that was presented horizontally. The same three factors that best explained year 1 data again best explained year 3 data.

Presented in Table 12 are item factor loadings for the Mathematics Computation portion of the test based on a three-factor solution. Six items changed factor assignment from year 1 to year 3. Items 17 and 18 returned to their original factor assignments in the third year. Item factor loadings for the Mathematics Concepts and Applications portion of the test are presented in Table 13. Seven of the 36 items changed factor assignments in year 2 and then returned to the same factor as in year 1 for year 3.

Confirmatory factor analysis (LISREL) was applied in order to determine whether the change in item loadings represented substantial changes in factor structures. The factor structure for the base year was used to constrain the pattern of fixed and free elements of Λ_x for subsequent years. Then goodness-of-fit indexes were compared. It was expected that the goodness-of-fit index for year 1 would be larger than that of year 2 or year 3. This expectation was met for Mathematics Computation but not for Mathematics Concepts and Applications. Goodness-of-fit indexes are presented in Table 14.

Factor analysis techniques were also used to examine skill-area subtest scores. It was conjectured that the use

Table 12

Grade 2 Mathematics Computation Factor Loadings of Items for 3 Years

Factor	Item No.	Item Loading*		
		1987	1988	1989
Addition: No Carrying	1	--	54	78
	2	83	80	58
	3	49	65	41
	4	73	81	67
	5	72	80	67
	6	66	55	45
	7	60	50	74
	8	63	67	58
	10	58	71	71
	14	--	67	61
	15	--	71	60
	17	--	53	--
	18	--	62	--
	20	--	41	51
Subtraction: No Regrouping	13	59	73	74
	14	56	--	--
	15	56	--	--
	17	47	--	45
	18	60	40	76
	19	63	58	62
	20	51	57	40
	21	72	63	68
	22	64	73	55
	23	85	56	57
	24	42	--	--
Addition-Subtraction: with Carrying and Regrouping	1	42	--	--
	9	86	91	87
	11	78	86	82
	12	82	84	89
	16	69	68	70
	24	72	67	66

*Decimal points omitted.

Table 13

Grade 2 Mathematics Concepts and Applications: Item Factor Loadings for 3 Years

Factor	Item No.	Item Loading*		
		1987	1988	1989
Geometry/Graphs Problem Solving	1	--	--	92
	3	--	82	--
	4	--	--	64
	5	57	--	74
	6	42	--	--
	7	--	86	--
	8	43	40	40
	10	60	--	--
	11	40	--	63
	12	61	--	46
	14	58	--	--
	15	--	49	--
	16	58	--	49
	20	--	65	--
	21	48	--	51
	22	--	69	--
	23	--	54	--
	24	47	--	65
	25	63	--	82
	26	46	--	67
	27	--	--	58
	29	44	--	--
	30	67	--	--
	31	--	76	--
	32	45	--	51
	34	66	--	61
	36	40	54	--
Place Value	1	44	--	--
	2	48	--	--
	3	82	--	--
	4	--	50	--
	5	--	59	--
	6	--	--	52
	7	88	--	--
	10	--	56	55
	12	--	40	--
	15	48	--	46

Table 13--continued.

Factor	Item No.	Item Loading		
		1987	1988	1989
Place Value (cont'd)	17	--	--	56
	18	--	--	46
	19	--	--	45
	21	--	40	--
	22	64	--	--
	23	50	--	--
	24	--	79	--
	25	--	65	--
	26	--	84	--
	29	--	--	47
	30	--	--	55
	31	81	--	--
	34	--	41	--
	35	55	--	--
Ordering/Time	1	--	57	--
	2	--	54	--
	3	--	--	80
	4	85	--	--
	7	--	--	82
	9	--	--	--
	11	--	63	--
	14	--	--	41
	16	--	58	--
	17	40	--	--
	18	40	--	--
	20	47	--	43
	22	--	--	73
	23	--	--	40
	27	43	--	--
	28	--	41	45
	29	--	40	--
	30	--	79	--
	31	--	--	76
	32	--	51	--
	35	--	--	46
	36	--	--	38

*Decimal points omitted.

Table 14

Grade 2 Goodness-of-Fit Index for Each Year and Subtest

Subtest	1987	1988	1989
Mathematics Computation	.94	.63	.65
Math Concepts and Applications	.90	.90	.90

of skill-area subtest scores would allow for a comparison of results without the use of dichotomous item data and tetrachoric correlation matrices. For all but two of the grade levels, there were too few subskills to result in stable estimates. Therefore, tables resulting from subskill analyses (excepting those for grades 3 and 5) are presented in the appendix.

Grade 3 Results

Exploratory factor analysis techniques on third-grade test items yielded results that indicated that the Mathematics Computation portion of the test could be explained by four factors in the base year. An examination of the items that loaded on each factor led to the conclusion that the factors are (a) addition/subtraction without carrying/regrouping, (b) subtraction with regrouping/ multiplication and division with remainders, (c) addition with carrying, and (d) multiplication facts/division facts.

An examination of scree plots and eigenvalue differences indicated that, in year 2, five factors could best explain the Mathematics Computation portion of the test. An examination of item factor loadings suggested that the factors are (a) addition/subtraction without regrouping, (b) subtraction with regrouping, (c) more complex multiplication/division with remainders, (d) addition with carrying, and (e) multiplication facts/division facts. The same five factors that explained year 2 data again best explained year 3 data.

Presented in Table 15 are item factor loadings for the Mathematics Computation portion of the test based on a four-factor solution. Twenty-five of the 44 items consistently loaded on the same factor for the 3 years. However, all of the items that had originally clustered as a fourth factor comprised of multiplication facts and division facts loaded on different factors in subsequent years.

Confirmatory factor analysis results (presented in Table 16) indicate that the base year's factor structure fits data from subsequent years almost as well. Nevertheless, the expected pattern of a lower goodness-of-fit index in the second year and then an increase in the index for the third year held.

Table 15

Grade 3 Mathematics Computation Item Factor Loadings for 3 Years based on Year 1

Factor	Item No.	Item Loading*		
		1987	1988	1989
Addition/ Subtraction with No Regrouping	1	40	--	--
	2	55	--	42
	3	--	45	58
	5	56	--	60
	6	58	--	53
	13	62	82	66
	14	49	68	76
	15	55	70	62
	16	66	73	69
	17	72	40	76
	18	64	51	55
	19	73	55	46
	20	65	58	59
	25	--	--	57
	26	--	--	81
	27	--	--	47
	28	--	--	56
	29	--	--	49
	38	--	--	49
Subtraction with Regrouping/ Multiplication and Division with Remainders	2	--	40	--
	21	46	--	--
	22	51	--	--
	23	55	--	--
	24	47	--	--
	27	--	49	--
	29	--	44	--
	30	--	56	57
	31	50	63	56
	32	64	67	64
	33	70	57	66
	34	67	53	76
	35	81	64	70
	36	68	63	71
	37	40	77	65
	38	--	70	--
	39	--	75	56

Table 15--continued.

Factor	Item No.	Item Loading		
		1987	1988	1989
Subtraction	40	--	71	53
with Regrouping/	41	45	70	59
Multiplication	42	47	66	43
and Division	43	69	57	51
with Remainders	44	59	62	56
(cont'd)				
Addition with	1	--	43	--
Carrying	3	52	--	--
	4	76	62	72
	5	--	69	--
	6	--	57	--
	7	79	64	75
	8	77	64	81
	9	84	62	82
	10	83	80	81
	11	73	68	72
	12	67	56	64
	25	--	52	--
	26	--	40	--
	28	--	47	--
Multiplication	15	--	--	41
Facts/Division	21	--	64	78
Facts	22	--	69	79
	23	--	79	82
	24	--	70	72
	25	80	--	--
	26	87	--	--
	27	60	--	--
	28	62	--	--
	29	69	--	--
	30	57	--	--
	37	69	--	--
	38	54	--	--
	39	40	--	--
	40	55	--	--

*Decimal points omitted.

Table 16

Grade 3 Goodness-of-Fit Index for Each Year and Subtest

Subtest	1987	1988	1989
Computation	.87	.83	.86
Concepts and Applications	.87	.83	.85

On the Mathematics Concepts and Applications portion of the test, four factors provide the best explanation. Factor loadings for the 3 years are presented in Table 17. The four factors are (a) ordering, patterning, and graphs, (b) problem solving, (c) geometry/measurement, and (d) algebra.

Nineteen of the 44 items consistently loaded on the same factor for the 3 years. Eight of the items loaded on the same factor in years 1 and 3 and on a different factor in year 2.

The goodness-of-fit indexes for the Concepts and Applications portion of the test are similar to those of the Computation portion. In years 2 and 3, goodness-of-fit indexes were lower than in year 1 (see Table 16). However, again, year 3 index is slightly higher than that of year 2.

The skill-area subtest scores factor loadings are presented in Table 18. It was determined that two factors best explain this portion of the test. They are (a) ordering/problem solving/geometry and (b) place value/

Table 17

Grade 3 Concepts and Applications Item Factor Loadings for 3 Years

Factor	Item No.	Item Loading*		
		1987	1988	1989
Ordering/Patterns Graphs	1	68	73	68
	2	64	73	64
	3	69	70	94
	4	--	--	98
	5	71	--	112
	7	65	60	--
	8	--	--	71
	9	63	--	69
	10	51	53	42
	11	64	63	--
	13	64	52	65
	15	--	43	57
	16	45	58	--
	17	43	45	51
	18	81	71	43
	19	73	75	--
	20	74	71	--
	22	40	--	40
	23	50	--	--
	24	73	50	--
	26	--	44	42
	29	47	--	--
	35	--	54	--
	38	--	--	48
	40	--	--	51
	41	--	--	52
	43	--	56	--
	44	--	78	45
Problem Solving	6	40	--	55
	7	--	--	46
	12	44	--	53
	14	47	58	59
	21	41	45	43
	23	--	62	--
	25	52	41	69
	26	63	--	--
	27	--	41	41
	28	57	66	70

Table 17--continued.

Factor	Item No.	Item Loading		
		1987	1988	1989
Problem Solving (cont'd)	29	--	47	48
	30	60	60	58
	31	60	63	84
	32	61	54	55
	33	71	48	90
	34	--	45	46
	35	52	--	40
	39	--	--	41
	41	56	--	--
	45	67	--	63
	46	61	61	64
	47	--	65	--
Geometry/ Measurement	5	--	67	--
	6	--	46	--
	8	--	67	--
	9	--	55	--
	16	--	--	42
	22	--	43	--
	23	--	--	73
	34	43	--	--
	38	43	46	--
	40	66	63	--
	41	--	52	--
	42	85	86	82
	43	59	--	40
	44	72	--	--
	45	--	45	--
	47	--	--	41
Algebra	4	47	56	--
	8	49	--	--
	11	--	--	40
	19	--	--	44
	36	85	90	91
	37	80	91	67

*Decimal points omitted.

fractions/measurement. Skill-area subtest 9 (measurement), made up of items 21, 29, 32, 34, 38, 39, 40, and 45, changed loadings from factor 2 to factor 1 and then back to factor 2 in years 1, 2, and 3, respectively. This pattern was not borne out in the factor analyses of dichotomous items.

Table 18

Grade 3 Mathematics Concepts and Applications Skill-Area Subtest Factor Loadings for 3 Years

Factor	Subtest No.	Subtest Loading		
		1987	1988	1989
Order/Problem Solving	2	80	75	67
Geometry	5	56	62	--
	7	61	72	78
	8	80	78	79
	9	--	60	--
Place Value/ Fractions/ Measurement	1	68	79	79
	3	56	54	58
	4	78	80	68
	5	--	--	68
	6	68	60	71
	9	64	--	65

Confirmatory factor analysis goodness-of-fit indexes for the Mathematics Concepts and Applications skill-area subtests exhibit similar patterns to those of Mathematics Computation for grade 3. Specifically, data in year 1 fit the specified factor structure slightly better than do years 2 and 3 data (see Table 19).

Table 19

Grade 3 Skill-Area Subtests Goodness-of-Fit Index for Each Year and Subtest

Subtest	1987	1988	1989
Concepts and Applications	.97	.96	.92

Grade 4 Results

Exploratory factor analysis of fourth-grade Mathematics Computation test items yielded results that indicated this portion of the test is best explained by two factors in year 1. An examination of the items that loaded on each factor led to the conclusion that the factors are (a) operations on whole numbers and (b) operation on fractions. An examination of scree plots and eigenvalue differences indicated that in years 2 and 3, this portion of the test was made up of the same two factors as in year 1.

Presented in Table 20 are item factor loadings for the Mathematics Computation portion of the test based on a two-factor solution. Forty-five of 50 items maintained consistent factor loadings over the 3-year period. One item loaded on the same factor in years 1 and 3 and a different factor in year 2.

Goodness-of-fit indexes resulting from confirmatory factor analyses (presented in Table 21) indicate that the base year's factor structure fits data from subsequent years well. The expected pattern of a lower goodness-of-fit index in the second year and then an increase in the index for the third year did not hold. Based on the fact that testing expectations for fourth grade varied from year 1 to year 3, the expectation was that factor structures would vary substantially from year to year.

On the Mathematics Concepts and Applications portion of the test, three factors provide the best explanation. The three factors are (a) geometry/measurement/graphs, (b) problem solving, and (c) algebra. Presented in Table 22 are item factor loadings based on a three-factor solution for each year. Twenty-two of 55 items loaded on the same factor for 3 years. Twelve of the 55 items loaded on the same factor in years 1 and 3 and a different factor in year 2 (Table 22).

Goodness-of-fit indexes resulting from confirmatory factor analysis (presented in Table 21) indicate that the

Table 20

Grade 4 Mathematics Computation Item Factor Loadings for 3 Years

Factor	Item No.	Item Loading*		
		1987	1988	1989
Operations on Whole Numbers and Decimals	1	55	56	43
	2	40	53	54
	4	70	49	65
	5	43	48	--
	6	69	48	61
	7	67	62	47
	8	46	66	54
	9	68	70	75
	10	66	63	59
	11	61	67	56
	12	68	74	70
	13	51	59	55
	15	50	45	62
	16	64	61	62
	18	50	--	55
	19	58	57	67
	20	77	68	65
	21	69	71	65
	22	49	66	40
	23	67	78	73
	25	58	45	42
	26	--	45	--
	27	50	48	40
	29	65	62	62
	30	58	56	48
	31	68	57	55
	32	65	69	64
	33	73	65	72
	34	66	62	70
	35	69	71	47
	36	69	72	56
	37	48	61	41
	38	61	57	--
	39	56	49	72
	41	--	45	--
	42	52	46	--
	44	69	55	55
	45	48	49	--
	48	47	53	--
	49	47	49	45

Table 20--continued.

Factor	Item No.	Item Loading		
		1987	1988	1989
Operations on Fractions	3	89	83	74
	14	97	84	73
	17	65	64	51
	28	82	80	68
	38	--	--	43
	40	92	82	84
	41	--	--	49
	42	--	--	55
	43	66	72	69
	45	--	--	50
	46	97	88	91
	47	81	70	84
	48	--	40	55
	50	84	84	79

*Decimal points omitted.

Table 21

Grade 4 Goodness-of-Fit Index for Each Year and Subtest

Subtest	1987	1988	1989
Computation	.89	.89	.88
Concepts and Applications	.89	.89	.88

Table 22

Grade 4 Mathematics Concepts and Applications Item Factor Loadings for 3 Years

Factor	Item No.	Item Loading*		
		1987	1988	1989
Geometry/	4	87	--	41
Measurement/	5	81	--	40
Graphs	10	--	59	--
	12	40	76	--
	13	58	--	--
	14	69	65	--
	15	50	70	--
	17	--	47	--
	19	63	52	--
	20	53	44	63
	21	46	58	67
	22	50	88	40
	23	--	80	43
	24	47	70	48
	25	--	--	40
	26	--	40	--
	27	59	--	40
	28	51	45	51
	29	48	58	65
	31	68	48	--
	34	--	--	55
	36	--	--	57
	37	--	--	58
	40	--	--	42
	41	41	--	--
	42	--	--	44
	43	55	--	40
	48	58	--	53
	49	44	--	63
	50	49	--	40
	51	--	49	56
	52	40	44	64
	53	40	--	67
	54	53	59	71
	55	46	--	57

Table 22--continued.

Factor	Item No.	Item Loading		
		1987	1988	1989
Problem Solving	1	--	--	69
	2	66	--	70
	3	51	43	59
	13	58	43	43
	17	46	--	--
	18	--	--	40
	23	43	--	--
	26	40	--	--
	27	--	40	--
	30	45	40	--
	32	47	40	42
	33	46	52	43
	34	46	50	--
	35	40	56	59
	36	54	40	--
	37	43	41	--
	38	57	60	49
	39	68	60	46
	40	43	46	--
	41	--	40	--
	42	64	--	--
	44	68	60	70
	45	39	55	--
	46	69	62	45
	47	56	60	55
	48	--	40	--
	49	--	47	--
	50	--	59	--
	53	--	45	--
	55	--	40	--
Algebra	1	--	58	--
	2	--	52	--
	4	--	63	--
	5	--	55	--
	6	69	--	71
	7	61	75	82
	8	74	43	66
	9	66	64	75
	10	74	--	71
	11	47	46	75
	12	--	--	57

Table 22--continued.

Factor	Item No.	Item Loading		
		1987	1988	1989
Algebra (cont'd)	14	--	--	77
	15	--	--	48
	16	41	61	68
	17	--	--	46
	18	--	59	--
	19	--	--	44
	25	40	41	--
	28	--	--	--
	30	--	--	40
	31	--	--	56
	45	--	--	40

*Decimal points omitted.

base year's factor structure fits data from subsequent years as well. Again, the expected pattern of a lower goodness-of-fit index in the second year and then an increase in the index for the third year did not hold.

Grade 5 Results

Exploratory factor analysis of fifth-grade Mathematics Computation test items yielded results that indicated this portion of the test is best explained by three factors. These factors are (a) operations on whole numbers, (b) operations on fractions, and (c) operations on decimal numbers. An examination of scree plots and eigenvalue differences indicated that in years 2 and 3 this portion of the test is best explained by the same three factors as in year 1.

Presented in Table 23 are item factor loadings for the Mathematics Computation portion of the test based on a three-factor solution. Twenty-six of the 55 items maintained consistent factor assignments over the 3-year period. Five items loaded on the same factor in years 1 and 3 and a different factor in year 2.

Goodness-of-fit indexes resulting from confirmatory factor analyses (presented in Table 24) indicate that the year 1 factor structure fits data for year 2 well, but the factor structure does not fit year 3 data as well. Expectations about testing did not change for fifth grade during the 3-year period, nor did average test scores differ.

On the Mathematics Concepts and Applications portion of the test, three factors provide the best explanation. The three factors are (a) number sentences/graphs, (b) problem solving/place value, (c) geometry/fractions (see Table 25). Thirty-four of 55 items loaded consistently on the same factors over the 3-year period. Two of 55 items loaded on the same factor in years 1 and 3 and a different factor in year 2.

Goodness-of-fit indexes resulting from confirmatory factor analyses (presented in Table 24) indicate that year 1 factor structure fits year 2 and year 3 data well. Since expectations about testing did not change nor did average test scores change substantially over the 3-year period,

Table 23

Grade 5 Mathematics Computation Item Factor Loadings for 3 Years

Factor	Item No.	Item Loading*		
		1987	1988	1989
Operations on Whole Numbers	1	--	--	40
	4	49	--	40
	7	67	--	--
	10	56	--	--
	11	62	53	47
	12	40	--	--
	17	49	--	--
	18	45	--	--
	19	53	47	51
	20	64	61	58
	21	56	52	57
	25	47	--	--
	27	42	--	--
	28	56	59	62
	29	59	--	--
	30	72	--	--
	31	51	--	51
	32	--	44	--
	33	44	--	--
	35	65	--	57
	37	66	46	52
	38	--	56	46
	39	41	--	--
	40	74	62	67
	41	64	56	61
	42	69	58	63
	43	--	49	58
	44	64	57	67
	45	--	60	51
	46	74	72	68
	48	78	78	73
	49	63	73	71
	50	--	66	65
Operations on Fractions	3	66	71	77
	6	80	71	77
	13	67	68	68
	14	80	78	77
	15	--	57	--

Table 23--continued.

Factor	Item No.	Item Loading		
		1987	1988	1989
Operations on Fractions (cont'd)	16	70	58	64
	18	--	40	--
	23	57	58	61
	26	62	63	56
	30	--	51	--
	32	57	--	58
	36	61	--	--
	38	48	--	--
	45	54	--	--
	50	59	--	--
Operations on Decimal Numbers	5	65	62	53
	7	--	--	41
	8	--	--	58
	10	--	--	61
	12	--	--	50
	15	63	--	56
	17	--	--	57
	18	--	--	55
	22	66	54	65
	24	79	67	66
	25	--	51	78
	27	--	--	53
	29	--	41	58
	30	--	--	55
	31	--	47	--
	33	--	43	--
	34	44	72	59
	35	--	51	--
	36	--	52	48
	39	--	44	46
	43	61	--	--

*Decimal points omitted.

Table 24

Grade 5 Goodness-of-Fit Index for Each Subtest

Subtest	1987	1988	1989
Computation	.89	.88	.65
Concepts and Applications	.90	.89	.88

it was not expected that factor structures should differ dramatically.

Factor analysis of skill-area subtest scores yielded results which indicated that the Mathematics Concepts and Applications portion of the test could be explained by two factors: (a) problem solving/graphs/geometry and (b) measurement (see Table 26). Skill area subtests 4, 6, 7, 8, and 9 changed factors on which they loaded over the 3-year period. Two of the 9 items that comprise skill-area 4 changed factor loadings when dichotomous items were factor analyzed. Four of the 6 items that comprise skill-area 6 changed factor loadings when dichotomous items were factor analyzed. Four of the 9 items that comprise skill areas 7, 8, and 9 changed factor loadings when dichotomous items were factor analyzed.

Goodness-of-fit indexes resulting from confirmatory factor analyses (presented in Table 27) on the Mathematics Concepts and Applications portion of the test for skill-area

Table 25

Grade 5 Mathematics Concepts and Applications Item Factor Loadings for 3 Years

Factor	Item No.	Item Loading*		
		1987	1988	1989
Number Sentences/ Graphs	1	40	80	--
	2	44	40	52
	3	66	70	51
	4	--	42	48
	5	60	54	--
	6	60	58	--
	7	42	50	40
	8	88	71	45
	9	59	53	72
	10	62	--	70
	11	72	58	--
	13	48	45	54
	18	74	89	81
	19	--	--	54
	20	67	58	--
	21	54	50	74
	22	61	68	60
	23	42	48	52
	24	77	56	55
	25	--	--	--
	26	--	45	41
	27	--	56	48
	28	--	56	48
	30	--	50	--
	32	--	62	40
	34	--	--	40
	37	51	40	--
	53	52	52	--
Problem Solving/ Place Value	12	51	60	45
	14	64	50	60
	15	--	44	--
	16	59	42	48
	17	64	64	77
	27	63	--	--
	28	42	--	--
	29	40	41	43
	30	46	--	--
	31	44	40	46

Table 25--continued.

Factor	Item No.	Item Loading		
		1987	1988	1989
Problem Solving/ Place Value (cont'd)	32	54	--	--
	33	40	40	63
	34	52	62	--
	35	58	52	47
	36	44	40	50
	37	--	--	40
	38	56	59	50
	40	50	55	65
	42	--	55	65
	43	74	61	50
	44	63	40	50
	45	58	65	--
	46	57	45	--
	50	--	56	--
	52	--	49	--
	54	--	40	--
Geometry/ Fractions	1	--	--	51
	4	56	--	--
	6	--	--	62
	11	--	--	42
	19	--	--	--
	25	--	40	40
	26	40	--	41
	39	48	--	51
	41	40	60	--
	42	45	--	--
	47	60	77	92
	48	73	78	85
	49	64	93	90
	50	40	--	51
	51	68	64	75
	52	49	--	--
	53	--	52	60
	54	49	--	40
	55	40	58	51

*Decimal points omitted.

Table 26

Grade 5 Mathematics Concepts and Applications Skill-Area Subtest Factor Loadings for 3 Years

Factor	Subtest No.	Subtest Loading*		
		1987	1988	1989
Problem Solving/ Graphs/Geometry	2	60	63	57
	3	65	58	64
	4	78	--	85
	5	49	71	60
	6	90	--	74
	7	--	67	--
	8	--	79	--
	9	--	76	--
Measurement	1	43	50	50
	4	--	76	--
	6	--	81	--
	7	91	--	87
	8	46	--	57
	9	73	--	57

*Decimal points omitted.

subtests indicate that the base year's factor structure fits data from subsequent years well. These results are similar to those of dichotomous item factor analyses.

Table 27

Grade 5 Goodness-of-Fit Index for Skill-Area Subtests

Subtest	1987	1988	1989
Concepts and Applications	.98	.98	.98

Alternative Considerations for Applications of Goodness-of-Fit Indexes

An argument could be made for applying factor structures of year 2 and year 3 to years 1 and 3 and years 1 and 2, respectively. If goodness-of-fit indexes follow the same pattern as those which resulted when the pattern of fixed and free elements of Λ_x was constrained by year 1 factor structures, then conclusions which are drawn are strengthened. These analyses were performed only on two grade levels. Second- and third-grade data were used to provide a contrast between grade levels that did and did not expect to test in the third focal year. Presented in Table 28 are goodness-of-fit indexes that resulted from confirmatory factor analysis dichotomous items constraining the pattern of fixed and free elements of Λ_x by year 2 factor structures.

Presented in Table 29 are goodness-of-fit indexes that resulted from confirmatory factor analysis of dichotomous items constraining the pattern of fixed and free elements of Λ_x by year 3 factor structures.

Table 28

Goodness-of-Fit Index Resulting from Application of Year 2 Factor Structures to Data from All Years

Grade	Subtest	Year on which y_2 factor structure was applied	Goodness-of-fit Index
2	Computation	1987	.62
		1988	.94
		1989	.63
	Concepts and Applications	1987	.92
		1988	.93
		1989	.91
3	Computation	1987	.84
		1988	.87
		1989	.84
	Concepts and Applications	1987	.82
		1988	.87
		1989	.83

Table 29

Goodness-of-Fit Index Resulting from Application of Year 3 Factor Structures to Data from All Years

Grade	Subtest	Year on which y_3 factor structure was applied	Goodness-of-fit Index
2	Computation	1987	.63
		1988	.63
		1989	.94
	Concepts and Applications	1987	.91
		1988	.92
		1989	.93
3	Computation	1987	.87
		1988	.86
		1989	.89
	Concepts and Applications	1987	.82
		1988	.83
		1989	.85

Patterns of goodness-of-fit indexes were similar when year 2 factor structures were applied to years 1 and 3 data to those that resulted when year 1 was used as the base year. The large differences in goodness-of-fit indexes between the base year and the other 2 years for second grade is manifested in this analysis as well. Third-grade analyses are also similar to those which used year 1 as the base year. Similar patterns are exhibited when year 3 factor structures become the "base year" by which all other factor structures are constrained.

Summary

Evidence was presented which suggests that test preparation activities varied from year to year. Beliefs about whether one's students would take the achievement test led to differences in classroom behavior with regard to test preparation activities.

There was little evidence to suggest that these test preparation activities are related to test scores. A few of the activities were significantly related to test scores, but different activities were significant at different grade levels.

Except at fifth grade, average Mathematics raw scores rose substantially from year 1 to year 2. Then, in year 3, average Mathematics raw scores remained essentially the same as in year 2. Average fifth-grade mathematics raw scores did not differ significantly for years 1, 2, and 3.

Factor analyses of dichotomous item responses provided evidence of some differences in factors onto which some items loaded over the 3-year period. However, except for second grade, confirmatory factor analysis results did not bear out substantial differences from year to year. Similar factor analysis techniques performed on skill-area subtests supported the results of dichotomous item response factor analyses for the most part. Constraining year 2 and year 3 factor structures on other years' data provided similar results to those when year 1 was considered the base year.

CHAPTER 5
DISCUSSION AND IMPLICATIONS FOR FURTHER RESEARCH

This study was designed to provide a description of the extent to which test preparation and curricular alignment activities took place over a 3-year period in a middle-sized metropolitan school district in north-central Florida. The focal years of the study encompassed the first year of the implementation of a new standardized achievement test through the third year of its use. The classes of teachers who taught the same grade level for the 3 years in question served as subjects. Data on both teacher behavior and student test performance were collected. Impact of teacher behavior on student performance at item, subskill area, and total test level were examined.

Summary of Findings

Evidence has been presented that indicates that teachers (and schools) change test preparation activities based on beliefs about whether students will be tested and, to some degree, on the perceived emphasis placed on testing by school boards. The greatest amount of test preparation activities took place in the second year of testing. Test preparation activities decreased in the third year at all grade levels.

This decrease could have been the result of a school board decision to cut back on the testing program. Although there were decreases in test preparation activities at all grade levels, the most dramatic cutbacks occurred in third and fourth grades. Teachers of these two grades did not expect to test their students in the spring of the third year. Teachers of the two grade levels (second and fifth grades) that expected to test their students in the spring of year 3 might have been expected to prepare their students for testing in the same ways as in the previous year or to continue increasing test preparation activities. However, it is presumed that the school board decision to cut back on testing carried with it an implied message that less emphasis would be placed on standardized test scores.

A two-way analysis of variance indicated a significant year-by-grade interaction on test preparation activities. Follow-up pairwise comparisons of grade levels within years led to the conclusion that beliefs about whether one's students would take the achievement test caused differences in classroom behavior with regard to test preparation activities and curricular alignment. Specifically, when teachers did not expect their students to be tested, they decreased test preparation activities.

Effects of Curricular Alignment, Test Preparation Activities on Math Test Scores

In order to determine whether test preparation activities had a significant impact on student test scores,

multiple regression techniques were employed. Class means served as the dependent variable. The results of these analyses seem to indicate that, for the most part, there is little relationship between test preparation activities and average scores. However, failure to find significance could be due to multicollinearity among the variables (see Appendix C). In view of this multicollinearity, it is also instructive to interpret correlation coefficients between test preparation activity scores and posttest means.

In the regression analysis, at second grade, test-taking skills activities were significantly related to test scores in year 2. An examination of the bivariate relationships for second grade reveals that test-taking skills activities were also significantly related to test scores in year 3. This analysis also indicated that motivational activities were significantly related to test scores in year 1.

Although no test preparation activities were significantly related to test scores for third grade in the regression analysis, an examination of bivariate relationships indicated that motivational activities were significantly related to test scores in year 1 and test-taking skills activities were related to test scores in year 2.

Motivational activities were significantly related to mathematics achievements for students in grade 4 for years 2 and 3 in the regression analysis. Since specific activities

were not dealt with separately, it is not possible to determine which kinds of motivational activities were most effective. Neither is it possible to determine if fourth-grade teachers used activities that differed from other grade levels or if the activities were simply most effective for fourth graders.

An interesting relationship emerged for year 1, fifth grade, where curricular alignment proved to be significantly related to test scores. That is, even though the previous year's test would no longer be used, teachers identified objectives that should receive greater emphasis during the subsequent school year. This more appropriate use of test information had a positive relationship with test scores.

Another surprising finding was that the only grade level in which a significant relationship between test scores and curricular alignment was evident in the regression analysis was at grade 2 during year 3. Teachers, principals, and district level personnel made a point of describing the various ways in which their curriculum was aligned with the standardized achievement test.

Average raw scores increased in the second year and then decreased in the third year at all grade levels. With the exception of third grade, the raw score differences seem insignificant. However, these relatively small raw score differences translate into larger differences in percentile ranks. (Standardized achievement test results are

traditionally reported in terms of percentiles for many school districts, including the district from which data for this study were drawn.) Table 30 illustrates this point. If teachers believe that test preparation activities are associated with an increase in a few raw score points, then these activities become understandable in light of their meaning in terms of percentile ranks.

It was expected that average raw scores would increase significantly the second year of test administration. This expectation was met at each grade level except grade 5. It is possible that, at grade 5, more emphasis is placed on statewide assessment objectives. This would be one explanation for the stability of fifth-grade scores over the 3-year period. The expectation that average raw score changes in year 3 would vary dependent upon whether or not testing was expected to occur was not met. Although average raw scores dropped at each grade level in year 3, none were significantly different than year 2 average raw scores. However, the transformation of these small raw score differences to percentile ranks (see Table 30) brought some concern to educators.

Factor Structures of Item Responses over a 3-Year Period

It is possible that the underlying factor structure of the test results changes from year to year. This possibility was addressed using both exploratory and confirmatory factor analysis techniques. Whereas Phillips

Table 30

Changes in Average Raw Score Means Expressed as Percentile Ranks

	Year		
Grade	1987	1988	1989
2	41	80	71
3	15	52	41
4	48	65	54
5	58	67	58

and Mehrens (1988) found that, for the most part, one factor adequately explained the test results of data they studied, accounting for at least 68% of the variance, all analyses in the present study produced results with more than one factor, all explaining less than 50% of the variance in each case. Additionally, the items did not always load onto the same factor over the 3-year period, suggesting that, in fact, factor structures fluctuate subtly from year to year.

Confirmatory factor analyses produced equivocal results. No statistical test of significance was used to compare the results year by year. Instead, the goodness-of-fit indexes were used to make year-by-year pairwise comparisons. Dichotomous item data as well as skill-area subtests were factor analyzed. The patterns of goodness-of-fit indexes for all grade levels and subtests were in agreement using the two

types of data. With the exception of second grade, Math Computation goodness-of-fit indexes did not differ substantially from year to year. However, the expectation of a better fit for the base year and a lower goodness-of-fit index in subsequent years held in each case. This pattern held when data for years 2 and 3 were used as the "base" year as well.

An examination of items that loaded on the different factors from one year to the next may shed some light on the issue of test preparation. For grade 2, Math Concepts and Applications, all of the items that called for students to solve problems that asked the question "how many in all?" changed factor loadings after year 1. The same is true for items that asked "how many are left?" These are words that are easily drilled and associated with operations of addition and subtraction, respectively.

For grade 3, Math Computation, multiplication facts and division facts (that is, multiplication of one digit by one digit and division by one digit with one-digit dividends) changed factors upon which they loaded after year 1. Again, these are items that are easily drilled previous to testing. If it is assumed that little or no curricular alignment took place in year 1 and (due to the fact that teachers did not expect to test third graders) in year 3 little or no curricular alignment occurred, the expectation would be that items that are most easily drilled change factor loadings in

year 2, and then in year 3 the factor structure would be more like that of year 1 when less preparation took place. This appears true for items that require students to count by 10s or round numbers or identify odd numbers and for number sentences asking for the replacement of a box with the correct number on the Concepts and Applications portion of the test.

The expectation for grade 4 factor analysis results was similar to that for grade 3. Three items, in particular, fit the expectation of changing factor loadings. Two of these items required students to solve addition of decimals in a horizontal presentation. The other item required the addition of decimals with carrying. Teachers could have familiarized students with this format and type of problem previous to testing.

On the grade 4 Math Concepts and Applications portion of the test, all items that require students to complete addition number sentences by replacing a box with the correct number changed factor loadings. It can be seen how easily teachers could incorporate this format into their curriculum prior to testing. Other items for which factor loadings were the same for years 1 and 3 but different for year 2 used the word "congruence" or a picture of a clock to test ability to tell time accurately. Again, these skills are easily drilled prior to testing.

Grade 5 items were expected to change factor loadings from year to year due to the fact that continuing test preparation activities occurred. The patterns and formats of items that changed factor loadings on the Math Computation portion of the test are not easily categorized. Nor can an explanation be made as to why these items changed factor loadings when similar items did not change. On the Concepts and Applications portion of the test, items dealing with congruence or geometry (pyramids, radius) changed factor loadings. These items could be remembered and drilled prior to testing.

For the most part, factor analyses of skill-area subtests supported the results of dichotomous item response factor analyses. Constraining year 2 and year 3 factor structures on other years' data provided similar results to those when year 1 was considered the base year.

Comparison of Results to Previous Research

This study differed from those of Mehrens and Phillips (1986, 1987, 1988) because it followed a single district's test results from the first year of the implementation of the testing program through its third year. This provides information regarding changes in curriculum and test preparation activities that Mehrens and Phillips could not account for. Mehrens and Phillips (1988) found that, generally, one factor accounted for at least 68% of variation and could adequately explain the test results. Results of

this study indicate that two or more factors are needed to explain the test results and only about 50% of the variation is accounted for.

Vojir and Shepard (1987), Cook, Eignor, and Taft (1988), and Miller and Linn (1988) studied the effects of varying curricula on high school test results. The present study focused on elementary grade levels. Shepard and Dougherty (1991) and Nolen, Haladyna, and Haas (1992) sought to define the types of test preparation activities used in classrooms and the extent to which these activities are used in classrooms. The purpose of the present study was to link information about amounts and types of test preparation activities, including curricular alignment, to effects of these activities on test results. Mehrens and Phillips (1986) sought to answer similar questions. However, their data focused on one year of test results, and definitions of varying curricula were based strictly on differences in textbooks and supplementary materials used. The present study was an attempt to overcome some of the limitations in the Mehrens and Phillips (1986) study. In terms of total test scores, conclusions reached in this study were similar to those of Mehrens and Phillips in that curricular alignment was found to have little impact on test results, even providing for more control over prior years' preparation activities. However, other test preparation activities were found to have some impact on test results. In terms of

factor structure, the results of this study differ from those of Mehrens and Phillips. In the present study, factor structures appear to change from year to year. Also, Mehrens and Phillips suggested that their test data could be adequately explained in terms of one major factor. Test results for the present study require more than one factor for adequate explanation.

Limitations

This study raises questions about test preparation activities and the effects on test factor structures; however, the sample consisted of students from only one school district. More extensive studies involving more school districts with similar results would provide for more definitive conclusions.

Although the teacher questionnaire that was used to ascertain levels of test preparation activities each year elicited a 100% response rate, each teacher answered only one questionnaire during the late winter/early spring of the third focal year of the study. It is believed that teachers provided truthful and candid answers; however, they were relying on their memories to report activities in their classrooms from the previous 2 years as well as the current school year. If this research study were to be replicated, it would be important to ascertain levels of test preparation activities each year.

No statistical tests of significance were used when confirmatory factor analysis techniques were applied. According to Kaplan (1990), several factors can affect the size of a chi-square statistic. Among those factors listed by Kaplan are "non-normality . . . and the sensitivity of the test to large samples" (pp. 140-141). Kaplan indicated that extreme caution should be applied to interpretation of chi-square tests for skewed data. A degree of skewness was observed in the present research study. This was apparent when it was observed that approximately 60% of items had item p-values greater than .80 at each grade level. Kaplan also indicated that large sample size interacts with misfitting models. These concerns should lead to caution in interpretation of the chi-square statistic.

Mulaik, James, Van Alstine, Bennett, Lind, and Stilwell (1989) discussed the use of the goodness-of-fit index. These researchers state that the goodness-of-fit index can be influenced by violations of assumptions and by model misfits. Though the goodness-of-fit index was used to compare the fit of data from one year to the next, interpretations should be made cautiously. Violations of the assumption regarding normality could have an effect on the goodness-of-fit index. The goodness-of-fit index was used with a baseline hypothesis that was not a no-factor null hypothesis. Mulaik et al. (1989) suggested this use while recommending that conclusions be drawn cautiously.

Implications for Further Research

An interesting replication of this type of study would be one in which test results of classrooms with teachers who engage in high levels of test preparation activities are compared to test results of classrooms with teachers who engage in low levels of test preparation activities. If test preparation activities are related to changes in test factor structures, the expectation would be that the test results of classes whose teachers provide for few test preparation activities would manifest little or no change in factors onto which items load over several years.

Many school districts use the same edition of a standardized achievement test for 5 to 7 years. Replicating this study for the inclusion of a longer period of time should also provide more definitive results.

This study provides a prototype for school districts to use in assessing the impact of test preparation activities on test scores. Circumstances under which testing schedules are changed midyear may not be possible or desirable. However, if teachers are surveyed yearly so that levels of test preparation activities can be reliably measured, similar analyses to those that were performed in this study will provide evidence of the extent of these activities and their impact on test scores as well as shifts in factor structures of test results.

APPENDIX A
TABLES RESULTING FROM FACTOR ANALYSIS OF SUBSKILL AREAS

Table A-1

Grade 2 Mathematics Computation Skill-Area Subtest Factor Loadings for 3 Years

Factor	Subtest No.	Subtest Loading*		
		1987	1988	1989
Addition/ Subtraction without Regrouping	1	58	66	--
	2	84	85	89
	4	76	76	77
Subtraction with Regrouping	1	--	--	63
	3	95	95	94

*Decimal points omitted.

Table A-2

Grade 2 Skill-Area Subtest Goodness-of-Fit Index for Each Year and Subtest

Subtest	1987	1988	1989
Mathematics Computation	.99	.93	.97
Math Concepts and Applications	.99	.98	.98

Table A-3

Grade 2 Mathematics Concepts and Applications Skill-Area Subtest Factor Loadings for 3 Years

Factor	Subtest No.	Subtest Loading*		
		1987	1988	1989
Problem Solving/ Measurement	1	72	71	66
	2	52	72	71
	3	84	75	84
	5	--	59	--
Geometry/Graphs	4	66	95	81
	5	80	--	68

*Decimal points omitted.

Table A-4

Grade 3 Mathematics Computation Skill-Area Subtest Factor Loadings for 3 Years

Factor	Subtest No.	Subtest Loading*		
		1987	1988	1989
Addition/ Subtraction	1	87	81	81
	2	83	88	89
	3	48	--	--
	4	59	79	80
Multiplication/ Division	6	82	84	81
	7	86	85	86
	8	62	63	57
Multiplication Facts	3	--	66	45
	5	91	60	74

*Decimal points omitted.

Table A-5

Grade 4 Mathematics Computation Skill-Area Subtest Factor Loadings for 3 Years

Factor	Subtest No.	Subtest Loading*		
		1987	1988	1989
Operations on Whole Numbers	1	65	57	73
	2	57	97	68
	5	79	--	76
	7	82	53	79
	8	81	--	71
Operations on Fractions	3	86	87	81
	6	86	90	82
Operations on Decimals	4	91	95	90
	5	--	48	--
	8	--	51	--

*Decimal points omitted.

Table A-6

Grade 4 Skill-Area Subtests Goodness-of-Fit Index for Each Year

Subtest	1987	1988	1989
Computation	.98	.95	.98
Concepts and Applications	.99	.33	.99

Table A-7

Grade 4 Mathematics Concepts and Applications Skill-Area Subtests Factor Loadings by Year

Factor	Subtest No.	Subtest Loading*		
		1987	1988	1989
Place Value/ Graphs/ Geometry	1	79	68	61
	4	47	--	71
	5	90	87	91
	6	57	72	--
Problem Solving/ Measurement	2	56	66	72
	3	99	89	86
	4	--	69	--
	6	--	--	72

*Decimal points omitted.

Table A-8

Grade 5 Mathematics Computation Skill-Area Factor Loadings for 3 Years

Factor	Item No.	Item Loading*		
		1987	1988	1989
Subtraction/ Division of Whole Numbers/Multiplication of Whole Numbers, Fractions	3	61	--	--
	6	60	57	--
	7	72	81	66
	8	69	57	60
	9	78	79	80
Addition, Subtraction of Fractions/ Subtraction of Decimals	2	71	64	72
	4	87	87	88
	5	49	59	--
Addition of Whole Numbers	1	85	87	79
	3	85	87	79
	5	--	--	62
	6	--	--	57

*Decimal points omitted.

Table A-9

Skill-Area Subtest Goodness-of-Fit Index Resulting from Application of Year 2 Factor Structures from Data of All Years

Grade	Subtest	Year onto which y_2 factor structure was applied	Goodness-of-fit Index
2	Computation	1987	.99
		1988	.99
		1989	.99
	Concepts and Applications	1987	.99
		1988	.99
		1989	.99
3	Computation	1987	.94
		1988	.97
		1989	.96
	Concepts and Applications	1987	.97
		1988	.98
		1989	.92

Table A-10

Skill-Area Subtest Goodness-of-Fit Index Resulting from Application of Year 3 Factor Structures from Data of All Years

Grade	Subtest	Year onto which y_3 factor structure was applied	Goodness-of-fit Index
2	Computation	1987	.99
		1988	.99
		1989	.99
	Concepts and Applications	1987	.99
		1988	.98
		1989	.99
3	Computation	1987	.92
		1988	.97
		1989	.98
	Concepts and Applications	1987	.94
		1988	.96
		1989	.97

APPENDIX B
FIRST 10 EIGENVALUES

Table B-1

First 10 Eigenvalues Resulting from Factor Analyses--1987

Grade	Subtest	Eigenvalues									
		1	2	3	4	5	6	7	8	9	10
2	Computation	9.0	2.0	1.3	0.7	0.6	0.6	0.5	0.4	0.4	0.3
3	Computation	15.6	3.1	2.3	2.1	1.5	1.4	1.3	1.0	0.8	0.7
4	Computation	18.4	3.9	1.7	1.6	1.5	1.2	1.1	1.0	0.9	0.7
5	Computation	15.9	2.5	2.0	1.5	1.4	1.3	1.1	1.0	0.9	0.8
2	Concepts & Application	9.7	2.3	2.0	1.4	1.3	1.1	1.0	0.8	0.7	0.6
3	Concepts & Application	16.7	3.0	2.0	1.8	1.4	1.4	1.2	1.1	0.9	0.8
4	Concepts & Application	21.3	2.3	1.7	1.3	1.3	1.2	1.1	1.0	0.9	0.8
5	Concepts & Application	19.3	2.6	2.0	1.3	1.3	1.1	1.1	1.0	0.9	0.8

Table B-2

First 10 Eigenvalues Resulting from Factor Analyses--1988

Grade	Subtest	Eigenvalues									
		1	2	3	4	5	6	7	8	9	10
2	Computation	10.2	3.0	1.4	1.0	0.5	0.5	0.5	0.4	0.3	0.2
3	Computation	15.9	3.2	2.6	2.0	1.7	1.2	1.2	1.1	0.9	0.8
4	Computation	16.7	4.1	2.0	1.5	1.2	1.2	1.1	1.0	0.9	0.8
5	Computation	14.7	2.4	2.0	2.0	1.1	1.1	1.0	1.0	0.9	0.8
2	Concepts & Application	10.8	1.7	1.5	1.4	1.3	0.9	0.9	0.9	0.7	0.6
3	Concepts & Application	19.8	2.5	2.2	2.0	1.3	1.2	1.1	1.1	1.0	0.9
4	Concepts & Application	22.1	2.1	2.0	1.3	1.3	1.2	1.1	1.0	0.9	0.8
5	Concepts & Application	19.2	2.3	2.1	1.4	1.3	1.2	1.0	1.0	0.9	0.8

Table B-3

First 10 Eigenvalues Resulting from Factor Analyses--1989

Grade	Subtest	Eigenvalues									
		1	2	3	4	5	6	7	8	9	10
2	Computation	9.6	3.0	1.1	0.8	0.7	0.7	0.6	0.4	0.4	0.3
3	Computation	15.7	2.8	2.7	2.0	1.5	1.5	1.2	0.9	0.8	0.7
4	Computation	18.2	3.1	2.1	1.7	1.5	1.3	1.1	0.9	0.9	0.8
5	Computation	17.8	2.2	2.0	1.4	1.2	1.1	1.0	0.8	0.8	0.7
2	Concepts & Application	11.5	2.0	1.6	1.4	0.8	0.8	0.8	0.8	0.6	0.5
3	Concepts & Application	16.9	3.1	2.1	1.9	1.3	1.3	1.1	1.1	1.0	0.8
4	Concepts & Application	20.6	2.3	2.0	1.4	1.4	1.2	1.0	0.9	0.9	0.8
5	Concepts & Application	19.6	2.6	2.0	1.4	1.3	1.2	1.1	1.1	0.9	0.8

APPENDIX C
CORRELATIONS AMONG TEST PREPARATION VARIABLES

Table C-1

Correlations for Grade 2, 1987

	MA	TTS	CA	SAP	Posttest
MA	1.00				
TTS	.54*	1.00			
CA	.85*	.37*	1.00		
SAP	.56*	.38*	.38*	1.00	
Posttest	.38*	.31	.26	.15	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-2

Correlations for Grade 2, 1988

	MA	TTS	CA	SAP	Pretest	Posttest
MA	1.00					
TTS	.54*	1.00				
CA	.63*	.41*	1.00			
SAP	.47*	.25*	.73*	1.00		
Pretest	.39*	.48*	.21	.23	1.00	
Posttest	.20	.53*	.23	.29	.27	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-3

Correlations for Grade 2, 1989

	MA	TTS	CA	SAP	Pretest	Posttest
MA	1.00					
TTS	.27	1.00				
CA	.51*	.41*	1.00			
SAP	.42*	.03	.42*	1.00		
Pretest	.32	.14	.05	.62*	1.00	
Posttest	.22	.43*	.29	.27	.70	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-4

Correlations for Grade 3, 1987

	MA	TTS	CA	SAP	Posttest
MA	1.00				
TTS	.20	1.00			
CA	.60*	.18	1.00		
SAP	.29	.51*	.14	1.00	
Posttest	.43*	.22	.30	.21	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-5

Correlations for Grade 3, 1988

	MA	TTS	CA	SAP	Pretest	Posttest
MA	1.00					
TTS	.70*	1.00				
CA	.49*	.69*	1.00			
SAP	.60*	.37	.42*	1.00		
Pretest	.12	.13	.19	.38*	1.00	
Posttest	.30	.46*	.34	.28	.71*	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-6

Correlations for Grade 3, 1989

	MA	TTS	CA	SAP	Pretest	Posttest
MA	1.00					
TTS	-.05	1.00				
CA	.12	.35	1.00			
SAP	.30	.30	.15	1.00		
Pretest	.11	.26	.27	.43*	1.00	
Posttest	-.08	.20	.23	.18	.73*	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-7

Correlations for Grade 4, 1987

	MA	TTS	CA	SAP	Posttest
MA	1.00				
TTS	.85*	1.00			
CA	.31	.02	1.00		
SAP	.43*	.57*	-.09	1.00	
Posttest	.27	.15	.31	.03	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-8

Correlations for Grade 4, 1988

	MA	TTS	CA	SAP	Pretest	Posttest
MA	1.00					
TTS	.66*	1.00				
CA	.43*	.51*	1.00			
SAP	.42*	.77*	.77*	1.00		
Pretest	.43*	.38*	.45*	.46*	1.00	
Posttest	.46*	.19	.28	.28	.81*	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-9

Correlations for Grade 4, 1989

	MA	TTS	CA	SAP	Pretest	Posttest
MA	1.00					
TTS	-.12	1.00				
CA	-.15	.22	1.00			
SAP	.05	.48*	.46*	1.00		
Pretest	-.29	.27	.59*	.36	1.00	
Posttest	-.37	.24	.02	.22	.56*	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-10

Correlations for Grade 5, 1987

	MA	TTS	CA	SAP	Posttest
MA	1.00				
TTS	.74*	1.00			
CA	.52*	.34	1.00		
SAP	.48*	.81*	.25	1.00	
Posttest	.31	.13	.68*	.01	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-11

Correlations for Grade 5, 1988

	MA	TTS	CA	SAP	Pretest	Posttest
MA	1.00					
TTS	.42*	1.00				
CA	.67*	.59*	1.00			
SAP	.43*	.47*	.58*	1.00		
Pretest	.30	.15	.43*	.16	1.00	
Posttest	.17	.23	.19	.01	.44*	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

Table C-12

Correlations for Grade 5, 1989

	MA	TTS	CA	SAP	Pretest	Posttest
MA	1.00					
TTS	.73*	1.00				
CA	.43*	.55*	1.00			
SAP	.41*	.60*	.46*	1.00		
Pretest	.22	.13	.30	-.17	1.00	
Posttest	.35	.40*	.43*	.03	.88*	1.00

*p<.05 (MA = motivational activities, TTS = test-taking skills, CA = curricular alignment, SAP = subject-area practice.

REFERENCES

Airasian, P. W. (1988). Measurement-driven instruction: A closer look. Educational Measurement: Issues and Practice, 7(4), 6-11.

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. Journal of Educational Measurement, 20(2), 103-118.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction of a test of scholastic aptitude. Journal of Educational Measurement, 10(2), 95-106.

Baker, E. L., & Herman, J. L. (1983). Task structure design: Beyond linkage. Journal of Educational Measurement, 20(2), 149-164.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test performance. Review of Educational Research, 53(4), 571-585.

Bracey, G. W. (1987, May). Measurement-driven instruction: Catchy phrase, dangerous practice. Phi Delta Kappan, pp. 683-686.

Brandt, R. (1989). On misuse of testing: A conversation with George Madaus. Educational Leadership, 46(7), 26-29.

Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. Journal of Educational Statistics, 13(3), 227-241.

Cannell, J. J. (1987). Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average. Daniels, WV: Friends for Education.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28(1), 61-75.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of effects of recency of instruction on the stability of IRT and conventional item parameter estimates. Journal of Educational Measurement, 25(1), 31-45.

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. The Elementary School Journal, 85(3), 315-336.

Ebel, R. L. (1965). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Flexer, R. J. (1991, April). Comparisons of student mathematics performance on standardized and alternative measures in high-stakes contexts. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39(3), 193-202.

Haertel, E., & Calfee, R. (1983). School achievement: Thinking about what to test. Journal of Educational Measurement, 20(2), 119-132.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991, June-July). Raising standardized achievement test scores and the origins of test score pollution. Educational Researcher, pp. 2-20.

Hall, B. W., Carroll, D., & Comer, C. B. (1988). Test use among classroom teachers and its relationship to teaching level and teaching practices. Applied Measurement in Education, 1(2), 145-156.

Houts, P. L. (1975, July/August). A conversation with Banesh Hoffman. National Elementary Principal, pp. 31-34.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing bias. Journal of Educational Measurement, 15(3), 209-225.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. Multivariate Behavioral Research, 25(2), 137-155.

Kellaghan, T., Madaus, G. F., & Airasian, P. W. (1982). The effects of standardized testing. Boston: Kluwer-Nijhoff.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Kulik, J. A., Kulik, C. C., & Bangert-Drowns, R. L. (1984). Effects of practice on aptitude and achievement test scores. American Educational Research Journal, 21(2), 435-447.

Linn, R. L. (1989, March). Has item response theory increased the validity of achievement test scores? Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." Educational Measurement: Issues and Practice, 9(4), 5-14.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18(2), 109-118.

Madaus, G. F. (Ed.). (1983). The courts, validity, and minimum competency testing. Boston: Kluwer-Nijhoff.

Madaus, G. F., Airasian, P. W., & Kellaghan, T. (1980). School effectiveness: A reassessment of the evidence. New York: McGraw-Hill.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in conformatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitless or fraudulent? Educational Measurement: Issues and Practice, 8(1), 14-22.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. Journal of Educational Measurement, 23(3), 185-196.

Mehrens, W. A., & Phillips, S. E. (1987). Sensitivity of item difficulty to curricular validity. Journal of Educational Measurement, 24(4), 357-370.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp.. 13-103). New York: American Council on Education and Macmillan.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations of instructional coverage. Journal of Educational Measurement, 25(3), 205-219.

Millman, J., & Paulk, W. (1969). How to take tests. New York: McGraw-Hill.

Moore, J.C., Schutz, R.E., & Baker, R.L. (1966). The application of a self-instructional technique to develop a test-taking strategy. American Educational Research Journal, 3(1), 13-17.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105(3), 430-445.

Muthen, B. (1988, December). Instructionally sensitive psychometrics: Applications of the Second International Mathematics Study. Los Angeles: University of California, Center for Research on Evaluation Standards and Student Testing.

Muthen, B. (1989). Using item-specific instructional information in achievement modeling. Psychometrika, 54(3), 385-396.

Muthen, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Applications of a new IRT-based detection technique to mathematics achievement test items. Journal of Educational Measurement, 28(1), 1-22.

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. Educational Measurement: Issues and Practice, 11(3), 9-15.

Phillips, G. W. (1990). The Lake Wobegon effect. Educational Measurement: Issues and Practice, 9(4), 3, 14.

Phillips, S. E., & Mehrens, W. A. (1987). Curricular differences and unidimensionality of achievement data: An exploratory analysis. Journal of Educational Measurement, 24(1), 1-16.

Phillips, S. E., & Mehrens, W. A. (1988). Effects of curricular differences on achievement test data at item and objective levels. Applied Measurement in Education, 1(1), 33-51.

Popham, W. J. (1987, May). The merits of measurement-driven instruction. Phi Delta Kappan, pp. 679-682.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4(3), 207-230.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9(4), 401-412.

Rudman, H. C. (1977, November). The standardized test flap. Phi Delta Kappan, pp. 179-185.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 49(2), 252-279.

Shepard, L. A. (1989). Why we need better assessments. Educational Leadership, 46(7), 4-9.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? Educational Measurement: Issues and Practice, 9(4), 15-22.

Shepard, L. A., & Dougherty, K. C. (1991, April). Effects of high-stakes testing on instruction. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Smith, M. L. (1991, June/July). Put to the test: The effects of external testing on teachers. Educational Researcher, pp. 8-11.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Vernon, P. E. (1954). Symposium on the effects of coaching and practice in intelligence tests. British Journal of Educational Psychology, 24, 57-63.

Vojir, C. P., & Shepard, L. A. (1987). Effect of instructional history on item response theory estimates derived from a mathematics achievement test. Unpublished manuscript.

Wahlstrom, M., & Boersma, F. J. (1968). The influence of test-wiseness upon achievement. Educational and Psychological Measurement, 28, 413-420.

Waterman, M. (1988, October 12). The risks of comparing NAEP results. Education Week, p. 22.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8(2), 125-145.

BIOGRAPHICAL SKETCH

The author was born and raised in Jacksonville, Florida, and attended Duval County public schools. Upon graduation from Robert E. Lee High School, she entered the University of Florida where she earned a Bachelor of Arts degree in 1975.

In 1979 she entered the graduate school of the University of Florida, College of Education, in Foundations of Education. She earned a Master of Arts in Education degree in 1982.

In 1988, the author again enrolled in the graduate school of the University of Florida, College of Education, in Foundations of Education. During this time, her experience included working as research associate for the School Board of Alachua County. This position allowed the writer to be in contact with other research and testing professionals throughout the state and to develop an interest in the questions addressed in this study.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Education.

Linda Crocker

Linda Crocker, Chair
Professor of Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Education.

Craig L. Frisby

Craig L. Frisby
Assistant Professor of Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Education.

M. David Miller

M. David Miller
Associate Professor of Foundations of Education

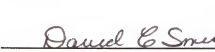
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Education.

R. Craig Wood
Professor of Educational Leadership

This dissertation was submitted to the Graduate Faculty of the College of Education and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Education.

May 1993

Chairman, Foundations of Education



Dean, College of Education

Dean, Graduate School